Testability and bounds in continuous instrumental variable models

Florian F Gunsilius*

Brown University

Job market paper. Comments welcome.
This version: November 6, 2018 First version: September 16, 2017.

Abstract

This article introduces two results for instrumental variable models with a continuous endogenous variable, allowing for the most general unobserved heterogeneity. The first result is a proof of a generalization of Pearl's conjecture (Pearl 1995*b*), showing that the exclusion restriction of an instrument cannot be tested in this setting without making structural assumptions. The proof is constructive and opens the door to potentially reestablish testability under weak assumptions. The second and main result is an approach for estimating sharp bounds on any possible causal effect, making no or minimal structural assumptions on the model besides the exclusion restriction. The key for this is to consider the instrumental variable model as two dependent stochastic processes and to construct an infinite dimensional linear program on their paths, the solution to which provides the counterfactual bounds. This framework can in principle encompass every structural assumption made in instrumental variable models and is the natural generalization of the complier-, defier-, never taker-, always taker distinction to the continuous setting. To showcase the estimation procedure, we obtain bounds on distributional causal effects of expenditures on leisure and food, using the 1996 UK Family Expenditure Survey. We find that food is a necessity while leisure is a luxury good, thereby corroborating the predictions from economic theory while introducing only minimal assumptions during the estimation process.

*For the latest version of this paper please click here. Correspondence: Florian_Gunsilius@brown.edu. I am very grateful to Susanne Schennach and Toru Kitagawa as well as Adam McCloskey and Ken Chay for their support and very helpful feedback throughout this project. I also want to thank Stefan Höderlein, Shakeeb Khan, Arthur Lewbel, Eric Renault, Jesse Shapiro, Michael Bedard, Simon Freyaldenhoven, Stefano Polloni, and Kevin Proulx, as well as the seminar participants at Brown University and Boston College for very helpful comments and discussions which helped me improve the article. All errors are mine. For the publication process, my goal is to separate the two main results and submit them as separate articles. In fact, the first result already exists as a separate note (Gunsilius 2018*b*).

1 Introduction

Estimating causal effects such as the average treatment effect (ATE) of an endogenous variable X on an outcome Y using a potential instrument Z for X is a fundamental problem in economics. The two main issues in this setting are (i) to check whether Z is exogenous to the model and, based on this, (ii) to estimate the respective causal effect from the data under minimal assumptions. The first issue calls for testable conditions under which Z satisfies the exclusion restriction. The second calls for a nonparametric framework for estimating causal effects which allows for nonlinearity and general unobserved heterogeneity, but does not require structural assumptions for the estimation process.

Despite their fundamental nature, there are only a few articles in the literature addressing these issues in this generality, all of which currently require a discrete X either in theory or in practice. This limits their applicability in many settings, as many endogenous variables of interest are continuous. For instance, one challenging problem is estimating a household's relative expenditure on different goods with total expenditure as the (continuous) endogenous variable (Blundell, Chen & Kristensen 2007; Imbens & Newey 2009). Currently, no practical nonparametric estimation procedures exist for this problem in our setting, where we avoid structural assumptions such as monotonicity or other shape restrictions of the involved production functions. Other general classes of applications stem from the evaluation of social programs with self-selection and randomized control trials with imperfect compliance. For instance, repayment obligations for micro-credit contracts, such as the interest rate, are sometimes chosen as the (continuous) treatment in randomized control trials in development economics and finance (Karlan & Zinman 2005). In these settings it frequently happens that the lending institution cannot perfectly comply with the assigned randomization of credit contracts due to internal guidelines, which induces an endogeneity problem (Karlan & Zinman 2011).

This paper deals with the case where X is continuous. We address both issues in this setting. First, we prove a slight generalization of Pearl's conjecture (Pearl 1995*b*), showing that without any structural assumptions the exclusion restriction of Z can in fact not be tested when X is continuous and hence needs to be an assumption in the model. Second, and this is the main result of this article, based on the exclusion restriction we provide a general framework for obtaining sharp upper and lower bounds on any causal effect of interest like the ATE, quantile effects, or distributional causal effects, in the case where X and possibly Y and Z are continuous. We do so under minimal assumptions and allowing for the most general unobservable heterogeneity, generalizing the results in Balke & Pearl (1994) and Balke & Pearl (1997). What is more, the framework we introduce is general enough to in principle encompass any structural form assumption made on the model, such as monotonicity. It hence constitutes an alternative to current approaches for phrasing questions in nonseparable triangular models and can be of use in basically any setting where the quantities of interest are causal relationships.

Both of the results we propose in this article possess properties which make them well-suited to

address issues (i) and (ii). As for the testability of the exclusion restriction, we want to emphasize that the proof of Pearl's conjecture is constructive. In particular, we can pinpoint when testability fails and hence provide a first step towards reestablishing testability in this general setting. As for our main result, one major contribution is that we have found a way to make the general theoretical framework for estimating bounds on causal effects applicable in practice, even in the general case where all variables are continuous. In particular, we introduce a general "sampling of paths" approach in order to approximate the optimal solutions of the infinite dimensional problems well in practice. This is the first instance in the literature where an estimator of this generality is feasible in the continuous case. As a result, applied researchers can also use our estimator for robustness checks prior to a parametric estimation, as our approach is laid out to handle general forms of randomness in the data.

In order to demonstrate the power of this approach, we obtain upper and lower bounds on distributional causal effects of expenditures on food and leisure using the 1995/1996 UK Family Expenditure survey (FES), only assuming that expenditure is continuous. We find strong evidence that leisure is a luxury while food is a necessity good. We do not make structural assumptions on the model other than continuity, so that we corroborate these economic postulates completely nonparametrically and without introducing further structural assumptions for the estimation procedure.

This article is structured as follows. In section 2 we provide the formal setting for both results and their connections to the literature. Section 3 contains the impossibility result on the testability of the exclusion restriction of Z. Section 4 contains the main results of this article on the estimation of causal effects. We outline the problem and the underlying intuition in subsection 4.1. We construct the infinite dimensional linear programs in the main subsection 4.2, discuss their theoretical properties in subsection 4.3 and their statistical properties in subsection 4.4. In section 5 we introduce our practical implementation of the infinite dimensional linear programs by our "sampling of paths"-algorithm. Section 6 contains our empirical application where we estimate consumer expenditure. Section 7 concludes. The appendix contains an overview of the mathematical notation we use as well as all proofs.

For readers with an applied focus we provide a running example about randomized control trials with imperfect compliance for a continuous treatment throughout subsections 4.1 and 4.2 to convey the main intuition of section 4. The end of each part of the running example is denoted by $\triangle$. One can get the fundamental idea of section 4 by only reading its introduction (page 10), the running example (pages 13, 16, and 19), the main result (Theorem 1, page 19), and the corollary for obtaining general causal effects like the ATE (Corollary 1, page 23), skipping everything else, in particular subsections 4.3 and 4.4.

2 Formal setting and relation to the literature

Throughout this article it will be convenient to have a structural representation of an instrumental variable model available:

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

Here, Y is the outcome variable, X is endogenous in the sense that it depends on the unobservable variable V, and Z is a potential instrument satisfying the relevance condition $Z \not\perp\!\!\!\perp X$, where "$\perp\!\!\!\perp$" denotes independence. Both V and U are unobserved random variables and can be of arbitrary and even of infinite dimension, just like Y, X, and Z. The production functions h and g are unknown.

Model (1) represents the most general structural form of an instrumental variable model (Pearl 1995*b*, Heckman 2001, Heckman & Vytlacil 2005 and references therein) and is general enough to encompass all important counterfactual relations, while being precise enough to enable all necessary mathematical derivations.[1] In the economics literature (1) is known as a *nonseparable triangular model* and has been the focus of a large number of articles, many of which deal with the identification of the production function h. Our focus in this article is on identifying and estimating causal effects of an exogenous shift of X on Y.

The main assumption on the error terms made in these models is the exclusion restriction $Z \perp\!\!\!\perp (V, U)$. It implies that the only influence the instrument Z has on the outcome Y is through X, i.e. $Z \perp\!\!\!\perp Y | X, (V, U)$. This follows from the fact that Z is not present in the second stage of model (1), which implies that there are no unobserved variables which jointly affect Z and Y; hence, if Z satisfies the exclusion restriction, then it is exogenous to the model. Also note that we require full independence of the instrument and not conditional mean independence or even weaker requirements. The reason is that we work in the most general nonlinear and nonparametric setting and do not want to make assumptions on the functional form of h or g. In the first part of this paper, i.e. section 3, we show that the exclusion restriction is not testable in the general setting with continuous X. From section 4 on we therefore assume it.

The results in this article are most closely related to two broad strands in the literature. First, and most generally, model (1) is the most comprehensive form of a nonseparable triangular model, which connects our results to this literature. Second, and more specifically, our results are connected to the literature on identification of causal- or treatment effects, often in the setting of program evaluation with self-selection (Imbens & Wooldridge 2009) or randomized control trials under imperfect compliance (Imbens & Angrist 1994, Angrist, Imbens & Rubin 1996). In the latter setting Y is the outcome, X is the actual treatment received, and Z is the original treatment assignment, the "intent-to-treat" variable, which serves as an instrument for the treatment X.

[1] Other representations of instrumental variable models are graphs (Pearl 1995*a*) and counterfactual notation (Rubin 1974). The appendix in Pearl (1995*b*) gives an overview of the connections between the different representations.

In the literature on general nonseparable triangular models there has been a surge in interest on the (partial-) identification of causal effects. Imbens & Newey (2009) derive identification results for the ATE and quantile effects while allowing for Y, X, and Z to be continuous, which makes their result closely related to ours in terms of the setting. In order to derive these results the authors require functional form assumptions on the relation between X and Z, however, as they are also interested in (point-) identifying h. In particular, they assume that g is strictly monotonic and continuous in Z and that U is univariate. Other identification results in nonseparable triangular models often focus on the production function h and also either 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, requiring the endogenous variable to be discrete. In contrast, our approach is completely general in that it does not require any *a priori* functional form assumptions, but provides a new way to include them in the model in the most general setting where X is allowed to be continuous.[2]

The article which is most closely related in the general theoretical handling of the problem is Chesher & Rosen (2017) which introduces a general framework for partial identification, using the Aumann integral and Artstein's inequality. Their framework works with a general model $\mathcal{M}$ based on general structural assumptions. Our approach is comparable in its generality even though we exclusively focus on causal effects. What distinguishes our approach from theirs is the applicability in practice. In fact, our identification and estimation strategy via infinite dimensional linear programs enables us to derive a practical and consistent estimation procedure for the most general setting, in addition to the general framework we set up. In contrast, results relying on Artstein's inequality run into severe curses of dimensionality for endogenous variables with many points in their support as the inequalities describing the identified region grow very rapidly (Beresteanu, Molchanov & Molinari 2012). Garen (1984) treats schooling as a continuous variable for similar reasons, for instance. Through our optimization approach we deal with this problem by a new "sampling of paths"-approach, which makes our problem applicable in the most general continuous setting in practice. Our approach also subsumes the discrete and binary approaches.

Two other important articles focusing on the partial identification in general models are Beresteanu, Molchanov & Molinari (2012) and Galichon & Henry (2011) which use the concept of random sets and optimal transport, in particular Choquet's capacity theorem (Choquet 1954), to obtain sharp bounds on different properties of interest. The former article assumes discrete X while the latter article deals with partially identifying unknown parameters in a structural setting. Their approaches, despite different from ours, are interesting as they rely on capacity theory. In fact, we write our problem as a *general capacity problem* (Anderson & Nash 1987, Anderson, Lewis & Wu 1989, Lai & Wu 1992) on a functional space in order to arrive at our

[2]Our framework also encompasses the case of discrete or binary X, in which case the stochastic processes we introduce would be point-processes.

partial identification result, which is a generalization of Choquet's capacity problem.

Recently, Mogstad, Santos & Torgovitsky (2018) introduced an optimization approach which deals with obtaining general causal effects in the setting where the endogenous variable is binary. Other articles relying on (finite) dimensional linear or convex optimization programs for estimating bounds on quantities of interest in general models are Chiburis (2010), Demuynck (2015), Honoré & Tamer (2006), Manski (2007), Molinari (2008), Lafférs (2015), Kamat (2017), Torgovitsky (2016), and Kitamura & Stoye (2018). Our approach is the first to define a general infinite dimensional linear program and enables the researcher to estimate causal effects in a general continuous setting with only minimal assumptions.

In the literature on the testability of the exclusion restriction Pearl (1995*b*) and Manski (2003) were the first to derive a testable implication for when Z satisfies the exclusion restriction when X is discrete. Extending these results, Kitagawa (2010) and Kitagawa (2015) derive a test when Y is continuous and X and Z are binary in different settings, also testing monotonicity of the instrument. Kitagawa (2009), again for binary X and Z, shows that Pearl's instrument inequality gives a sharp testable implication allowing Y to have arbitrary support. Kédagni & Mourifié (2015) show the same conclusion for binary Y and augment Pearl's inequalities by further inequalities in the case where Z has more than two points in its support. We complement these results by proving Pearl's conjecture (Pearl 1995*b*), showing that the exclusion restriction is not testable when X is continuous. The way we set up the proof of Pearl's conjecture is also related to recent results from the literature on identification of nonseparable triangular models such as Hoderlein, Holzmann, Kasy & Meister (2017) and Hoderlein, Holzmann & Meister (2017) which show that (point-) identification in general nonseparable triangular models with general heterogeneity is not achievable without relatively strong structural assumptions.

The most closely related results in spirit for tackling issue (ii), despite the fact that they focus on the case where Y, X, and Z are binary, are Balke & Pearl (1994) and Balke & Pearl (1997) which provide tight upper and lower bounds on any potential causal effect of interest under no or minimal structural assumptions. 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 X is binary. Kitagawa (2009) derives closed-form solutions for sharp bounds on causal effects for a continuous Y and binary X and Z, building on ideas from both Balke & Pearl (1997) and Manski (1990). Recently, Russell (2017) derived sharp bounds on causal effects for Y, X discrete using Artstein's inequality and optimal transport theory similar to Galichon & Henry (2011). We complement these results by obtaining sharp bounds in the most general continuous setting, in particular generalizing Balke & Pearl (1994) and Balke & Pearl (1997) to the continuous setting.

The focus of our main result is on identification and estimation, but we do derive large sample asymptotics for the infinite dimensional linear programs, for each bound separately. These asymptotic results might be of interest in themselves as they constitute new results for nonparametric plug-in estimators of infinite dimensional constrained linear programs.[3] To obtain accurate cov-

[3]The proofs are general enough to also hold for nonparametric plug-in estimators of many infinite dimensional

erage of the identified set based on our large-sample result, one can use established results from the literature such as Imbens & Manski (2004).

3 First result: Testability of the exclusion restriction of Z

Here we prove a generalization of Pearl's conjecture (Pearl 1995*b*), showing that the exclusion restriction $Z \perp\!\!\!\perp (V, U)$ is not testable when X is continuous. In fact, we prove a slightly stronger result: we show that the exclusion restriction on Z is not testable even if $g(z, U)$ is assumed to be invertible in U in model (1). In addition, we show that the conjecture holds even if the instrument Z is allowed to have (finitely many) atoms, which covers every probability measure encountered in practice. Our method of proof does not cover the case where the respective measure has a countably infinite number of atoms[4], but those measures cannot be recovered in practice via standard statistical methods and are therefore pathological. Allowing for g to be invertible in U allows us to make a connection to the voter paradox in majority voting. In fact, the key for proving the stronger version of the conjecture is to construct a Condorcet cycle in uncountable state space.

Our definition of a continuous random variable is the most general in that we consider continuous random variables to be *nonatomic*. A nonatomic probability measure P_X is one where for every measurable set A in the Borel σ-algebra $\mathscr{A}_X$ with $P_X(A) > 0$, there exists $B \in \mathscr{A}_X$ with $B \subset A$ and $P_X(A) > P_X(B) > 0$. For example, probability measures which are absolutely continuous with respect to Lebesgue measure are nonatomic, but there exist many nonatomic measures which do not possess a density with respect to Lebesgue measure. If the respective σ-algebra is the Borel σ-algebra, then a measure is nonatomic if and only if every point has measure zero. In the following, calligraphic letters like $\mathcal{Y}$, $\mathcal{X}$, and $\mathcal{Z}$ denote the support of the measures P_Y, P_X, and P_Z, respectively.[5]

We now state Pearl's conjecture (Pearl 1995*b*) in our notation and give some intuition. The formal proof is relegated to the appendix.[6] In our set-up we can state Pearl's conjecture as well as its slight extension as follows.

Conjecture (Pearl (1995*b*))**.** *Let Y, X, and Z be random variables, where P_Z is a general probability measure with at most finitely many atoms. If the marginal $P_{X|Z=z}$ of $P_{Y,X|Z=z}$ is nonatomic for almost every $z \in \mathcal{Z}$, then $P_{Y,X|Z=z}$ can be generated through model* (1)*:*

$$\begin{aligned} y &= h(x, v) \\ x &= g(z, u) \qquad \textit{with } Z \perp\!\!\!\perp (V, U). \end{aligned}$$

constrained convex programs under minimal modifications.

[4] I thank Susanne Schennach for this remark.

[5] Note that the supports can be of arbitrary dimension in principle as long as they are Polish, i.e. separable, complete, and metrizable spaces.

[6] Also see the note Gunsilius (2018*b*) for this result.

This even holds if the function $g(z,u)$ is assumed to be invertible in U.

To see that this conjecture implies that the exclusion restriction for Z cannot be tested when X is continuous, notice that if every possible observable measure $P_{Y,X|Z}$ can be generated by the model whilst assuming $Z \perp\!\!\!\perp (V,U)$, then there can be no testable restrictions on the model. That is, if a model manages to fit every possible data generating process, then it can never be tested, as there are no settings in which it can fail to explain the observables.

Let us give two important remarks on this conjecture. First, note that the above statement is slightly more technical than the wording of Pearl's original conjecture; in particular, Pearl simply stated that "if the variable X is continuous, then every joint density $f_{Y,X|Z=z}$ can be generated by model (1)". Since we work in an instrumental variable model, the important probability measures are $P_{Y,X|Z=z}$ and $P_{X|Z=z}$, so that the continuity assumption needs to be upheld with respect to the conditional measure $P_{X|Z=z}$ and not the unconditional measure P_X. In fact, this is what Pearl meant when he stated the conjecture, as the proof of the conjecture relies on a Lemma in Pearl (1995*b*) (Lemma 1 below), which explicitly relies on the fact that the measure $P_{X|Z=z}$ is nonatomic almost everywhere—Pearl himself exclusively worked with density functions and hence implicitly assumed that all relevant distributions, especially $P_{X|Z=z}$ and $P_{Y,X|Z=z}$, are absolutely continuous with respect to Lebesgue measure, a stronger condition than we uphold.

Second, the above statement of the conjecture is more general in that it allows for the assumption that g is invertible in U, which is a structural assumption in this model. Also, we only require $P_{X|Z=z}$ to be nonatomic for almost every $z \in \mathcal{Z}$ and only require that Z has at most finitely many atoms. The other conditional distributions including Y such as $P_{Y,X|Z=z}$ are left completely unspecified.

Let us now give the idea for the proof of the conjecture. For this we need to introduce the concept of *generators*, a term coined in Pearl (1995*b*).

Definition 1 (Generator)**.** *Given a probability measure $P_{X|Z}$, a function $x = g(z,u)$ is a* generator *of $P_{X|Z}$ if and only if there exists some probability measure on the domain of U such that $g(z,U)$ is distributed as $P_{X|Z=z}$. A generator is one-to-one if and only if for almost every $z_i, z_j \in \mathcal{Z}$ and $u \in \mathcal{U}$ $g(z_i,u) = g(z_j,u)$ implies $z_i = z_j$.*

We need generators because of the following lemma, which is the key for proving the conjecture. The special case of this lemma for probability density functions was derived and proved in Pearl (1995*b*).

Lemma 1. *Any probability measure $P_{Y,X|Z}$ whose marginal $P_{X|Z}$ has a one-to-one generator can be generated by* (1)*.*

This lemma reduces the problem of proving the conjecture to simply proving that there exists a one-to-one generator for each possible $P_{X|Z}$. Note that by using Lemma 1 we do not make any assumptions on the distribution of Y, so that we can allow for general distributions here too,

which does not change the result. In fact, the whole proof works with the properties of $P_{X|Z=z}$, which we assume to be nonatomic for almost all $z \in \mathcal{Z}$.

Now the main idea for the proof of the conjecture lies in the realization that a *one-to-one* generator is a special type of production function $g(z,u)$ mapping the unobservable U onto X for almost every z. For illustrative purposes of this point, let X, Z, and U be discrete with three values each: x_1, x_2, x_3 each with probability $\frac{1}{3}$, z_1, z_2, and z_3 each with probability $\frac{1}{3}$, and the same for U. Then we can build a matrix where the m^{th} row represents $f_{z_m}(U)$ and the n^{th} column represents x_n. Each cell (m,n) of the matrix contains the index $i \in \{1,2,3\}$ of u_i assigned to (m,n) by the generator $g(z_m,\cdot)$. For example the matrix could look like this:

$$\begin{array}{c} \\ z_1 \\ z_2 \\ z_3 \end{array} \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \underbrace{\begin{pmatrix} \mathbf{u_3} & u_2 & u_1 \\ \mathbf{u_3} & u_1 & u_2 \\ u_1 & u_2 & u_3 \end{pmatrix}}_{\text{not one-to-one in } z} \end{array}$$

Consider the cell $(2,1)$. The entry u_3 means that the value u_3 gets mapped to x_1 for z_2. In this case, the map $g(z,u)$ represented by this matrix is clearly not a one-to-one generator as needed, since it maps u_3 onto x_1 for both z_1 and z_2.

In order for $g(z,u)$ to be a one-to-one generator, one simply needs to guarantee that no column contains two or more equal numbers. However, in our generalization of Pearl's conjecture we also assume that $g(z,u)$ is a measure preserving isomorphism in *both* variables. We must hence require that neither columns nor rows contain two equal numbers. This requirement is analogous to constructing a Condorcet cycle from voter theory, which is possible for $n = m$:

$$\begin{array}{c} \\ z_1 \\ z_2 \\ z_3 \end{array} \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \underbrace{\begin{pmatrix} u_1 & u_2 & u_3 \\ u_2 & u_3 & u_1 \\ u_3 & u_1 & u_2 \end{pmatrix}}_{\substack{\textbf{Condorcet cycle:} \\ \text{invertible in both } z \text{ and } u}} \end{array}$$

Therefore, our proof of the conjecture proceeds by showing that a Condorcet cycle can always be constructed in the continuous setting, which we do in the formal proof in the appendix.

Three further remarks about the conjecture and our proof are in order. First, in order for the construction of the Condorcet cycle to work in uncountable state space, a nonatomic $P_{X|Z=z}$ for almost all z is crucial. In fact, every Borel set $E \subset [0,1]$ of positive measure contains an uncountable number of elements, so that even if there is an uncountable number of z, one can always find a permutation of E such that almost every $x \in E$ corresponds to exactly one z, hence

constructing a Condorcet cycle.

Second, our proof of Pearl's conjecture also gives a simple explanation for why the exclusion restriction of Z can be testable when X is discrete. As an example, let X and Z each take three values, x_1, x_2, and x_3 as well as z_1, z_2, z_3, and let U be uniformly distributed on the unit interval. Moreover, assume that $P_{X|Z=z_1}(x_1) + P_{X|Z=z_2}(x_1) > 1$. In this case, there cannot exist a Condorcet cycle. In fact, no matter how we partition the unit interval for U, the fact that $P_{X|Z=z_1}(x_1) + P_{X|Z=z_2}(x_1) > 1$ always implies that there is some Borel set $E_u \subset [0,1]$ of measure

$$P_U(E_u) = P_{X|Z=z_1}(x_1) + P_{X|Z=z_2}(x_1) - 1 = \varepsilon_u > 0,$$

which gets mapped to x_1 for z_1 and z_2, no matter the measure preserving map $g(z,u)$, implying that there cannot be a Condorcet cycle or even a one-to-one generator as depicted in the illustration of the proof.

Third, note that we can intuitively understand the construction of the Condorcet cycle in our proof as a special construction of the function $g(z,u)$, which has a non-standard form. In the literature on nonseparable triangular models, one usually assumes monotonicty and continuity of g. In contrast, for our non-testability result we need to allow for very general classes of functions g so that we can always replicate the observable joint distribution $F_{Y,X|Z=z}$. In this sense, it might be possible to reestablish testability of the exclusion restriction even in the continuous setting by restricting the set of allowable functions g.[7]

Overall, the fact that we managed to prove the impossibility theorem constructively makes us hopeful that we can find more appropriate functional form assumptions on the production functions to make the exclusion restriction testable in the continuous setting, too. In the general setting of model (1), this section shows, however, that we need to assume the exclusion restriction. Let us now turn to the main result of this article, the estimation of causal effects.

4 Main result: Obtaining bounds on causal effects

This is the main section of this paper where we provide theoretical and practical results to obtain tight upper and lower bounds on the counterfactual probability $P_{Y|X^*}$ which in turn yields bounds on the ATE or any other possible causal effect.[8]

Let us quickly lay out the problem of identifying causal effects in the model

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

[7] I thank Toru Kitagawa for this remark.

[8] Throughout, we will denote the counterfactual probability for exogenous X by $P_{Y|X^*}$. This notation has been used in other articles dealing with the identification of nonseparable triangular models, see for instance Torgovitsky (2015).

where $Z \perp\!\!\!\perp W$ for $W \coloneqq (V, U)$. This is an equivalent form of (1), which makes it easier to get the main point across in this subsection. Intuitively, the fundamental problem for deriving bounds on the ATE or quantile effects, or any causal effect for that matter, is to identify the counterfactual probability $P_{Y|X^*}$ for an *exogenous* change in X. Note that $P_{Y|X^*}$ is unobservable and does not coincide with the observable $P_{Y|X}$, since the latter gives the probability of the outcome given an *endogenous* change in X. This is because X depends on W so that we need to represent the exogenous version as

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

using model (1), where $\mathcal{W}$ is the support of P_W and where $P_{Y|X,W=w}$ is a conditional measure. Unfortunately, the only distribution in the data which gives us correct information on the DGP is $P_{Y,X|Z}$, because X is endogenous, so that the observable $P_{Y|X}$ is different from $P_{Y|X^*}$. To see this, write $P_{Y,X|Z}$ in our model as

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

as $Z \perp\!\!\!\perp Y|X, W$ (Balke & Pearl 1994, p. 50).

Now, if X were actually exogenous we would have $P_{X|Z,W} = P_{X|Z}$, so that we could identify $P_{Y|X^*}$ by

$$P_{Y|X} = \frac{P_{Y,X|Z}}{P_{X|Z}} = \frac{\int_{\mathcal{W}} P_{Y|X,W=w} P_{X|Z,W=w} P_W(dw)}{P_{X|Z}} = \frac{P_{X|Z} \int_{\mathcal{W}} P_{Y|X,W=w} P_W(dw)}{P_{X|Z}} = P_{Y|X^*}.$$

That is, the observable $P_{Y|X}$ would coincide with $P_{Y|X^*}$ in the case where X is actually exogenous. As soon as X is endogenous, the above line of reasoning does not work anymore, so that $P_{Y|X^*} \neq P_{Y|X}$ in general. Under endogeneity of X we will only be able to identify bounds on $P_{Y|X^*}$ without further assumptions. For this it is convenient to model W as two latent variables U and V as in model (1). The endogeneity of X can then be captured by the fact that V depends on U, so that one needs to replace $P_W(dw)$ by the joint measure $\mu(dv, du)$.

4.1 Outline of the problem, assumptions, and intuition

In this subsection we give an outline for why the problem is so difficult to solve for continuous X by contrasting it to the case where Y, X, and Z are discrete, in particular binary. Throughout, we need to uphold the exclusion restriction, based on our impossibility result from the previous section.

Assumption 1 (Exclusion restriction)**.** *In model* (1) *it holds that* $Z \perp\!\!\!\perp (V, U)$.

The main challenge in the continuous setting is that under continuous X one needs to distinguish between an infinite number of production functions g and h, whereas in the binary case

there are only four different functions (the standard defier, complier, never taker, and always taker distinction). Let us be more precise.

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 U and V, as in Balke & Pearl (1994). The important thing to realize is that u and v *index* the respective production functions $g(z, \cdot)$ and $h(x, \cdot)$ from model (1). In this case, both U and V possess four realizations as there are four possible functions mapping Z to X and X to Y in each case: u_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), u_2 corresponds to the function g mapping $Z = 0$ to $X = 0$ and $Z = 1$ to $X = 1$ (the compliers), u_3 corresponds to the function g mapping $Z = 0$ to $X = 1$ and $Z = 1$ to $X = 0$ (the defiers), and u_4 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 $v_1, \ldots, v_4$ in terms of realizations of Y and X.

Generalizing this idea, it is not hard to see that the cardinality of $\mathcal{U}$ is of the rate n^m, where m is the number of points in $\mathcal{Z}$ and n is the number of points in $\mathcal{X}$.[9] Analogously for v, which is of the rate q^n, where q is the number of points in $\mathcal{Y}$. Therefore, already a simple generalization to the 3-values case will, without further assumptions, lead to $3^3 = 27$ values for both v and u. Cheng & Small (2006) provide an identification result in this setting by circumventing this issue in the 3-value problem: they make monotonicity assumptions on the production functions g and h in order to bring down the cardinality of v and u from 27 to 4 again.

Now for general continuous Y, X, and Z the cardinality of U and V will be that of the power set of the continuum, $2^{\mathfrak{c}}$. Therefore, the natural σ-algebra for the measure spaces on which U and V are defined is the Lebesgue σ-algebra which has the cardinality of the power set of the continuum. We assume that Y, X, and Z lie in the standard Borel σ-algebra. This set-up is the direct analogue of the finite dimensional setting, as the cardinality of the Borel σ-algebra is that of the continuum.

The key realization then is that $g(z, u)$ and $h(x, v)$ induce stochastic processes described by the conditional measures $P_{X|Z=z}$ and $P_{Y|X^*=x}$, respectively. To see this recall that a stochastic process X_t with $t \in T$ for an arbitrary index set T is a collection of T random variables $X : (\Omega, \mathscr{S}, P) \to (\mathbb{R}, \mathscr{B}, P_X)$, where P_X is the pushforward measure of P through X. In mathematical terms the random variable is a function between Ω and $\mathbb{R}$ such that $P_X(E) = P(X^{-1}E)$ for every $E \in \mathscr{B}$.

Now compare this to our nonseparable triangular model (1). By definition, the two measures $P_{X|Z=z}$ and $P_{Y|X^*=x}$ are the pushforward measures of P_U and P_V via the production functions $g(z, U)$ and $h(x, V)$ for almost all z and x, respectively. In particular, $(\mathcal{U}, \mathscr{A}, P_U)$ and $(\mathcal{V}, \mathscr{A}, P_V)$ take the place of $(\Omega, \mathscr{S}, P)$ and $g(z, \cdot)$ as well as $h(x, \cdot)$ take the place of X_t. One can therefore view $P_{X|Z}$ and $P_{Y|X^*}$ as being induced by the stochastic processes $X_z(u)$ and $Y_x(v)$ which in turn are induced by $g(z, U)$ and $h(x, V)$ pushing forward P_U and P_V, respectively. In the binary case,

[9] As in the previous section, calligraphic letters define sets or the spaces on which the corresponding random variables are defined.

this means that each type, e.g. the never taker, is one path of the 2-point stochastic process. In our more general setting, the never taker corresponds to the path which is constant at 0 for every realization z.

Therefore, $P_{Y|X^*=x_0}(E_y)$ for some realization x_0 of X^* and some event E_y is the probability that a path of $Y_x(v)$ goes through the set E_y at the point x_0. Analogously, the observable $P_{Y,X|Z=z}(A_{y,x})$ defines the probability that both processes $Y_x(v)$ and $X_z(u)$ go through the set $A_{y,x}$ at z. Upper and lower bounds on $P_{Y|X^*=x_0}(E_y)$ for fixed x_0 and E_y can then be obtained by maximizing or minimizing the probability that the path of the stochastic process $Y_x(v)$ goes through E_y at x_0 under the constraint that the probability that both processes $Y_x(v)$ and $X_z(u)$ jointly go through the set $A_{y,x}$ at z is $P_{Y,X|Z=z}(A_{y,x})$ for each set $A_{y,x}$ and realization z of Z.

Running example (1). To make our theoretical results more tangible we consider the following setting. Suppose we have data from a (fictive) randomized control trial estimating the efficacy of repayment obligations in microcredit contracts on general outcomes like business profits or expenditure, a complex problem in development finance (Karlan & Zinman 2005; Karlan & Zinman 2011; Field, Pande, Papp & Rigol 2013). The (continuous) treatment X is a certain repayment obligation of the microcredit contract, e.g. the interest rate, the grace period in days, or the overall financing period—let us say it is the interest rate. The outcome Y could be weekly business profits, household income, or the savings rate over the next several years (Field, Pande, Papp & Rigol 2013). Since those randomized control trials rely on a cooperating lender, usually a bank, there can be compliance issues. In particular, the bank might give out credits with different conditions than those initially randomly assigned in order to not risk losing money (Karlan & Zinman 2011). This results in an endogeneity problem as the adjustments made by the bank are usually based on proprietary information and are not observed by the economist. The instrument in this case is the initially randomly assigned interest rate. The goal for us in this setting is to still obtain sharp upper and lower bounds on different causal effects of interest.

With a continuous treatment the amount of the treatment (i.e. the interest rate in percent) actually received can deviate from the assigned treatment in an infinite number of ways. The experimenter can only work with the overall distribution $P_{Y,X|Z=z}$ since she knows that the observed distributions $P_{Y|X}$ and $P_{X|Z}$ are for an endogenous X due to the imperfect compliance. Our approach therefore consists of obtaining the "best"- and "worst" case scenario for the counterfactual probability $P_{Y|X^*=x_0}(E_y)$ for some event E_y we are interested in given an *exogenous* assignment of an interest rate of x_0 percent.[10] Assume we are interested in the event E_y that the overall savings rate after 5 years increased by 10% given an assigned interest rate of x_0 percent.

The idea of our approach is to introduce a theoretical device: hypothetical participants, analogous to the binary case, i.e. the never takers, always takers, compliers and defiers for the relation

[10]Note that we can also work with more general events E_x for X^*, not just a fixed number x_0. For example, we can condition on the fact that X^* lies in a range of percentages. In addition, we can identify other causal effect like the ATE, but we stick with the counterfactual probability to convey the intuition for this part of the running example.

between X and Z. In a continuous setting there are an infinite number of such hypothetical participants, not just four. Each of those participants—and this is the main idea—is specified by a path of the stochastic processes X_z and Y_x of the first and second stage of the instrumental variable model, i.e. "response profiles", which are indexed by the unobservable U and V, respectively. A response profile for X_z tells the respective participant what interest rate x she actually receives for each assigned interest rate z. For instance, the never taker is the path of the stochastic process which is always zero for each value of Z, i.e. this participant's response profile is to receive, for every z the value zero (i.e. the bank will not give her any credit, no matter what interest rate was initially assigned to her). The path of the complier would be the 45° line, since this participant perfectly obtains whichever condition was randomly assigned (this is a participant with a very good credit score: no matter what interest rate we assign randomly, the bank will give the participant exactly this interest rate).

Analogously a response profile $Y_x(v)$ "tells" a participant how much she reacts (y) to a certain level of the treatment x. In this case the process Y_x models the fact that participants respond differently in terms of savings- or investment rate to different levels of interest rates; for instance, the hypothetical participant who does not respond to the treatment at all has the path $Y_x \equiv 0$ for any value of x; she would never invest anything, no matter how good the conditions of the interest rate. Very intuitively, the idea then is to find the *relative number* of hypothetical participants (i.e. the *weights* on the respective paths of the stochastic processes) that *maximizes* (for an upper bound) or *minimizes* (for a lower bound) the relative number of hypothetical participants which respond to a given interest rate of x_0 percent by an increase in their overall savings rate after 5 years by 10% *subject to the constraint* that the composition of the hypothetical participants *replicates* the joint observable *weight* (i.e. probability) $P_{Y,X|Z=z}$.

Put slightly differently, for an upper (lower) bound we want to see how many (few) hypothetical participants *who respond exactly an increase of the* 5 *year savings rate by* 10% *for an assigned interest rate of* x_0 *percent* we can possibly fit in our model such that we still obtain an overall composition of hypothetical participants which replicate the observable composition induced by $P_{Y,X|Z=z}$ for every z. This framework is the natural generalization of the seminal distinction into always taker, never taker, complier, and defier from Angrist, Imbens & Rubin (1996) to the continuous setting.

Note in this respect that Z need not be a continuous random variable, which it rarely is in applied research. We allow for Z to be discrete or binary; the corresponding stochastic process X_z in the latter case would then be a 2-point process instead of a continuous stochastic process. In fact, we can actually allow for all Y, X, and Z to be discrete; this case is easier to handle than the general continuous case as we then have fewer (i.e. a finite number) of stochastic processes to optimize over. In fact, our framework also works in the case where Y and X are discrete or binary, too, which makes our approach a direct generalization of Balke & Pearl (1994) and Balke & Pearl (1997). In the following our benchmark set-up will be for continuous Y, X, and Z since this is the most general and complicated setting. △

The most general assumption we need in this respect is to ensure that those probability measures actually define stochastic processes. The basic assumption we make throughout is

Assumption 2 (Regularity)**.** *(i) Y, X, Z, U, and V are random variables taking values in* $[0,1]$.

Under Assumption 2 the induced stochastic processes are defined on $[0,1]^{[0,1]} \subset \mathbb{R}^{[0,1]}$ with the canonical σ-algebra induced by their respective cylinder sets. For example, the cylinder sets for $X_z(u)$ are

$$\{X_z(u) \in [0,1]^{[0,1]} : (X_{z_1}, \dots, X_{z_k}) \in B, B \in \mathscr{B}_{[0,1]}\}.$$

Therefore, each u and v indexes one *equivalence class* of modifications of stochastic processes defined by the cylindrical σ-Borel algebra and the respective measure.[11]

As a second remark on this assumption, note that requiring the random variables to take values in $[0,1]$ is not restrictive as there always exists a measure preserving isomorphism from any Polish space equipped with the Borel σ-algebra and some nonatomic measure P onto the unit interval equipped with Lebesgue measure (Bogachev 2007, Theorem 9.1.11). Since we will make use of the compactness of $[0,1]$ in some proofs, our results hold for *compact* subsets of Polish spaces.[12]

As a third and final remark, note that we set out to solve a more general problem than other identification results in nonseparable triangular models (e.g. Imbens & Newey 2009, Torgovitsky 2015 and references therein), where assumptions on the dimension of U and V are constantly made. Our focus lies on partially identifying $P_{Y|X^*}$ and *not* point-identifying the production function h, so that we do not make any functional form assumptions on h or g which would lead to further structural assumptions on the dimensionality of $\mathcal{V}$ and $\mathcal{U}$. All we need for our setting is a measure space which has *cardinality* $2^{\mathfrak{c}}$ in order to index all possible paths of the respective stochastic process $Y_x(v)$. The unit interval equipped with the Lebesgue σ-algebra is the easiest set-up to work with for our purposes and this is why we choose it.[13]

We can now use Kolmogorov's Extension Theorem (Bauer 1996, Theorem 35.3) in order to prove that the stochastic processes $Y_x(v)$ and $X_z(u)$ induced by the measures $P_{Y|X^*=x}$ and $P_{X|Z=z}$ exist.

Proposition 1 (Existence of the stochastic processes)**.** *Under Assumption 2 the measures $P_{X|Z=z}$ and $P_{Y|X^*=x}$ induce stochastic processes X_z and Y_x.*

Note that right now, Proposition 1 only guarantees that there exist measure spaces $(\Omega, \mathcal{A}, P)$ on which X_z and $(\widetilde{\Omega}, \mathcal{A}, \widetilde{P})$ on which Y_x are defined. In the next section we construct measure preserving isomorphisms between the laws of the stochastic processes $P_{Y|X^*=x}$ and $P_{X|Z=z}$ and

[11] A modification of a path $X_z(u)$ of a stochastic process is another path $X'_z(u)$ which differs from $X_z(u)$ only on a set of measure zero.

[12] It is most likely possible to extend the results to the non-compact setting under formal complications.

[13] It is also from this point of view that a general consideration of the case of a continuous X is important, as it allows us to circumvent the combinatorial issues which often plague approaches using discretization arguments.

the unit interval, so that we can define X_z on $([0,1], \mathcal{A}_{[0,1]}, P_U)$ and Y_x on $([0,1], \mathcal{A}_{[0,1]}, P_V)$ and in turn interpret $P_{Y|X^*=x}$ and $P_{X|Z=z}$ as the laws of the stochastic processes defined by model (1) which are indexed by v and u, respectively. We will already use U and V in this section, while the formal construction is in the next section.

In principle, Assumptions 1 and 2 are the only assumptions we need to make our subsequent approach work. One can therefore identify the ATE without any structural assumptions on g and h. However, the σ-algebra induced by the cylinder sets on $\mathbb{R}^{[0,1]}$ is rather coarse. In particular, the space $C([0,1])$ is not measurable in this σ-algebra (Bauer 1996, Corollary 38.5). This technicality is important as in many applications it is warranted to make the assumption that the paths of the stochastic processes satisfy certain properties, like continuity, monotonicity, or some other general (nonparametric) functional form assumption. For instance in section 6 we apply our result to estimate Engel curves, for which the continuity assumption is natural.

Running example (2). Throughout this article, the only assumption we make is that the stochastic processes (i.e. response profiles of the hypothetical participants) $Y_x(v)$ and $X_z(u)$ are continuous, i.e. to require that $Y_x(v)$ and $X_z(u)$ lie in $C([0,1])$ instead of $\mathbb{R}^{[0,1]}$. This assumption translates to requiring that g and h be continuous with respect to X and Z in model (1), respectively. In fact, continuity of h with respect to X means that small changes in the administered treatment X do not lead to vastly different outcomes Y. Continuity of g with respect to Z means that if the actual interest rate Z intended to be assigned only changes slightly, then the actual assigned interest rate X does not vastly change. Our approach also works without this assumption, allowing for jumps. For instance in the microcredit example it can happen that the credit is not granted even though a certain interest was randomly assigned, which would yield jumps down of the stochastic processes.

This also gives a first indication of how we are able to include general (nonparametric) structural assumptions in model (1): we translate the respective assumption from the production functions $g(z,u)$ and $h(x,v)$ to the response profiles $X_z(u)$ and $Y_x(v)$. Different assumptions will rule out or allow for different profiles. For instance, in the microcredit randomized trial it is reasonable to allow for jumps in the response profiles: it might happen that the bank gives a credit to a person that was not assigned to receive a credit, in which case there would be a jump in the interest rate for this person. The same holds in the other direction: the bank might decide to not give a credit to some person even though it was instructed to do so, in which case there would be a jump to zero in the interest rate.

Since assumptions like continuity, monotonicity, or certain starting points of the processes rule out a number of paths of the stochastic processes, they will lead to tighter bounds *ceteris paribus* when upheld in the model, something which has also been described before (Heckman & Vytlacil 2005). Our general framework allows for an intuitive explanation: we do not have to account for as many potential strategies when optimizing, which in general leads to tighter bounds. Finally note that the generality of our approach provides a framework for potentially testing the strength

of these structural assumptions: assumptions which yield tighter bounds are stronger in general. $\triangle$

Continuity assumptions are structural assumptions on the model, but are very weak; in fact, they don't even have a counterpart when X is binary. Moreover, the space $C([0,1])$ is a Polish space, i.e. a completely metrizable and separable metric space, so that one can construct an indexing of the respective paths of a stochastic process by u and v rather easily by relying on results in Kuratowski (1934) which introduces a practical way to construct a measure preserving isomorphism from some Polish space onto the unit interval. In fact, we use Kuratowski's construction when setting up the indexing of the stochastic processes for our main result in the next section.

Another way to restrict the paths of the stochastic processes would be to assume that the stochastic processes $Y_x(v)$ and $X_z(u)$ induced by $P_{Y|X^*=x}$ and $P_{X|Z=z}$ lie in the Skorokhod space $D([0,1])$ of functions which are right-continuous and possess left limits (Billingsley 1999, Chapter 3). In that case, one would relax the continuity assumption to allow for a countable set of jump-discontinuities. The Skorokhod space when equipped with the corresponding Skorokhod metric is a Polish space, so that it would be equally convenient to obtain an indexing of the paths of the respective stochastic processes by u and v as in the space $C([0,1])$, using the general construction in Kuratowski (1934).

As mentioned, we will stick with the continuity assumption throughout as we also assume continuity of Engel curves in our application. The most well-known and basic condition for obtaining smoothness of the paths of the stochastic processes is the following.

Assumption 3 (Hölder continuous paths)**.** *There are fixed real numbers α, β, γ, $\delta > 0$ and real fixed finite constants c_x, $c_y > 0$ such that*

$$E(|Y_{x_1} - Y_{x_2}|^\alpha) \coloneqq \int_{[0,1]} |Y_{x_1}(v) - Y_{x_2}(v)|^\alpha \pi_1\mu(dv) \leq c_y|x_1 - x_2|^{1+\beta} \quad \textit{and}$$
$$E(|X_{z_1} - X_{z_2}|^\gamma) \coloneqq \int_{[0,1]} |X_{z_1}(u) - X_{z_2}(u)|^\gamma \pi_2\mu(du) \leq c_x|z_1 - z_2|^{1+\delta}$$

for all $x_1, x_2 \in [0,1]$ and $z_1, z_2 \in [0,1]$, and where $\pi_1\mu$ and $\pi_2\mu$ are the marginal measures of μ.

Two remarks about this assumption are in order. First, note that it actually implies that the paths of the stochastic processes are Hölder continuous, so that the corresponding stochastic processes can be defined on $C([0,1]) \subset \mathbb{R}^{[0,1]}$.

Proposition 2. *Under Assumptions 2 and 3 the stochastic processes $Y_x(v)$ and $X_z(u)$ induced by the disintegrated measures $P_{Y|X^*=x}$ and $P_{X|Z=z}$ possess a Hölder continuous modification of all orders $\lambda_Y \in (0, \frac{\beta}{\alpha})$ and $\lambda_X \in (0, \frac{\delta}{\gamma})$ and can hence be defined on $C([0,1])$.*

The proof of this proposition follows directly from Kolmogorov's Continuity Theorem (Bauer 1996, Theorem 39.3), and the fact that the continuous paths are actually Hölder continuous

follows from Theorem 39.4 in Bauer (1996). In fact, we chose this assumption, because it is the standard assumption in the literature for guaranteeing that the stochastic processes have a continuous modification.

Second, the actual choice of the variables α, β, γ, $\delta > 0$ as well as the constants c_x, $c_y > 0$ should be based on two things in principle: the *a priori* belief about the smoothness of the production functions $h(x, v)$ and $g(z, u)$ and the observable $P_{Y,X|Z=z}$, as Proposition 2 gives an *upper* bound on the degree of smoothness of the paths. In fact, as λ_Y and λ_X approach 0 the respective paths become less and less smooth and become only bounded in the limit. This limit, i.e. $\lambda_X = 0$ or $\lambda_Y = 0$ is excluded, however, as one cannot define the processes on $C([0, 1])$ in this case.

We also require a similar assumption on the observable distribution $F_{Y,X|Z=z}$, which we need in order to prove the large sample properties of the program in section 4.4.

Assumption 4 (Regularity of the observable distribution function)**.** *The observable measure $P_{Y,X|Z=z}$ induces a distribution function $F_{Y,X|Z=z}(y, x) \in C([0, 1]^3)$ which satisfies the following condition:*

$$\int_{[0,1]^2} \int_{[0,1]^2} |(s_y, s_x) - (t_y, t_x)|^{\eta_1} dF_{Y,X|Z=z_1}(s_y, s_x) dF_{Y,X|Z=z_2}(t_y, t_x) \leq c_{y,x} |z_1 - z_2|^{1+\eta_2} \qquad (2)$$

for some constants $\eta_1, \eta_2, c_{y,x} > 0$.

Assumption 4 is similar to Assumption 3. In particular, the regularity condition (2) is a sufficient condition for the constraint of our infinite dimensional linear program to be non-empty as we prove in Lemma 4 below. The non-emptiness of the constraint is important for us as it enables us to use a general form of the functional Delta-method in order to derive the large sample distribution for a natural nonparametric plug-in estimator. Both assumptions are very weak and should always be satisfied in models. If not, then one should work with more general stochastic processes anyways. Moreover, note that Assumptions 3 and 4 are only of theoretical relevance for us; in fact, when estimating the model in practice in section 5, we construct continuous stochastic processes by simply simulating paths from a trinomial tree, so that we do not have to deal with setting the appropriate constants from Assumptions 3 and 4 for this purpose.

Let us now turn to the main subsection of this article where we introduce the infinite dimensional linear programs which yield the bounds on our causal effects.

4.2 The formal result: Constructing the infinite dimensional linear programs

To make our main result easier to understand, let us state the main relations here again. Recall that we started out with unobservable measures $P_{Y|X^*}$ and $P_{X|Z}$, which induce stochastic processes on $[0, 1]^{[0,1]}$ by the construction as product measures as in Proposition 1. Proposition 2 shows that those product measures can be defined on $C([0, 1])$ under the continuity assumptions. Using the construction by Hess, which we state below in Lemmas 2 and 3, we define an indexing

of the paths of the stochastic processes $Y_x(v)$ and $X_z(u)$ for almost every v and u. This enables us to replace the general measure spaces $(\Omega, \mathscr{A}, P_V)$ and $(\Omega, \mathscr{A}, P_U)$ of stochastic processes by $([0,1], \mathscr{A}_{[0,1]}, \pi_1\mu)$ and $([0,1], \mathscr{A}_{[0,1]}, \pi_2\mu)$, which is important as it allows us to find the optimal joint measure μ on $[0,1]^2$ instead of some general abstract measure space. Using these ideas we can state the main result of this article.

Theorem 1 (Infinite dimensional linear program for bounds on counterfactual probabilities)**.** *Let Assumptions 1 – 3 hold. Then (lower/upper) bounds on the respective counterfactual distributions $F_{Y|X^*=x_0}(y^*)$ for fixed $y^* \in [0,1]$ and $x_0 \in [0,1]$ can be obtained as solutions to the following infinite dimensional linear programs:*

$$\begin{aligned} &\text{minimize/maximize}_{\mu \in \mathcal{P}^*([0,1]^2)} \quad \int_{[0,1]} H_{Y,X}(y^*, x_0, v)\pi_1\mu(dv) \\ &s.t.\ F_{Y,X|Z=z}(y,x) = \int_{[0,1]^2} G_{Y,X}(y,x,z,v,u)\mu(dv,du) \end{aligned} \tag{3}$$

for all $(y,x) \in [0,1]^2$ and almost all $z \in [0,1]$.[14] *Here, $\mathcal{P}^*([0,1]^2)$ denotes the set of all measures on $([0,1]^2, \mathscr{A}_{[0,1]^2})$ whose marginal measures $\pi_1\mu$ and $\pi_2\mu$ satisfy Assumption 3, i.e.*

$$\mathcal{P}^*([0,1]^2) \coloneqq \{\mu \in \mathcal{P}([0,1]^2) : Y_x \sharp \pi_1\mu \text{ and } X_z \sharp \pi_2\mu \text{ satisfy Assumption 3}\}. \tag{4}$$

$H_{Y,X}$ and $G_{Y,X}$ are integral kernels of the form

$$H_{Y,X}(y,x,v) = \mathbb{1}_{[0,y]}\{Y_x(v)\} \quad \text{and} \tag{5}$$

$$G_{Y,X}(y,x,z,v,u) = \mathbb{1}_{[0,y]\times[0,x]}\left\{Y_{X_z(u)}(v)\right\} \tag{6}$$

where $Y_x \sharp \pi_1\mu$ and $X_z \sharp \pi_2\mu$ denote the respective pushforward measures, i.e. laws of Y_x and X_z.

Let us give an intuitive explanation of this result in terms of our

Running example (3). The intuitive idea for Theorem 1 is as follows. Recall that the underlying idea is to work with the hypothetical participants, or more specifically: their response profiles which are defined by paths of the stochastic processes Y_x and X_z, respectively. We then obtain the counterfactual probability $P_{Y|X^*=x_0}(E_y)$ for some event E_y at exogenously assigned x_0 by changing the composition of response profiles in the model, as mentioned previously. Here "changing the composition" means to find a *joint* measure (i.e. weight) μ on the paths of the stochastic processes $Y_x(v)$ and $X_z(u)$.

The indexing of the paths $Y_x(v)$ and $X_z(u)$ needed for Theorem 1 is provided in Lemmas 2 and 3 below. This is *one way* of assigning an index u and v to (almost) every path of the stochastic processes X_z and Y_x, respectively, so that we are allowed to write $X_z(u)$ and $Y_x(v)$. This indexing has the additional benefit that we can work with the measure μ on the unit square $[0,1]^2$, which

[14] Note that throughout we define conditional distributions as $F_{Y,X|Z=z}(y,x)$ and do not write $F_{Y,X|Z=z}(y,x,z)$.

is the space where the indices of the response profiles U and V live. In this respect, it is also helpful to interpret μ as a probability on the response profiles instead of a weight—in this case one has one hypothetical participant who chooses between all admissible responses and puts a certain probability on these. The only assumptions we make in this setting is the exclusion restriction (Assumption 1), the normalization that all variables lie in the unit interval (Assumption 2), and the requirement that the paths of the stochastic processes are continuous (Assumption 3)—the approach works equally well when replacing the continuity assumption by some more general assumption, like allowing for jumps, however.

Also notice that Theorem 1 remains unchanged for binary or discrete instruments Z. All one has to do is change the stochastic processes X_z, but the optimization problem stays exactly the same. In fact, the case where all Y, X, and Z are continuous is by far the most complicated case since one then needs to optimize over the greatest number of stochastic processes. For binary Z, we obtain a 2-point stochastic process for continuous X. As mentioned, our optimization problem also works when all variables are discrete or binary.

Theorem 1 uses the probability interpretation in conjunction with the construction of the indices u and v of the paths of the stochastic processes; in particular, it tries to adjust the probability on the strategies of the hypothetical participant such that it maximizes (for an upper bound) or minimizes (for a lower bound) the probability of the hypothetical participant reacting with a 5 year savings rate of 10% to an assigned interest rate of x_0 percent such that the overall probability of the participant's response profile for both processes Y_x and X_z perfectly replicates the observable joint probability measure $P_{Y,X|Z=z}$. In this respect, the constraint set $\mathcal{P}^*([0,1]^2)$ contains all the admissible strategies, i.e. all structural assumptions we make on the stochastic processes. For instance, under the continuity assumption (Assumption 3) it contains only measures which induce continuous stochastic processes. Different functional form assumptions on the response profiles will change the elements $\mathcal{P}^*([0,1]^2)$, but do not change the structure of the linear programs. This makes including such assumptions convenient in our setting.

The notation in (3) does make clear that there is indeed a need for optimization. To see this, recall that we allow for a large class of stochastic processes in general, i.e. $h(X,v)$ and $g(Z,u)$ are not injective in v and u, which translates to the fact that the stochastic processes $Y_x(v)$ and $X_z(u)$ each have crossing paths. If paths cross at a point, it is not clear which response profile induced this action, so that knowing the joint distribution $P_{Y,X|Z=z}$ does not guarantee that we know the distributions $P_{Y|X^*}$ and $P_{X|Z}$. In fact, there are infinitely many probability measures μ on the paths of the processes Y_x and X_z which satisfy the constraint as $F_{Y,X|Z=z}(y,x)$ only gives a joint measure for both paths *simultaneously*, but it does not pin down paths individually as the processes are dependent: Y_x depends on the position of X_z. Note in this respect that the exclusion restriction $Z \perp\!\!\!\perp (V,U)$ is incorporated in the function $G_{Y,X}$; in fact, the processes $Y_x(v)$ and $X_z(u)$ are constructed independently, so that the only thing that makes them dependent is the fact that $Y_x(v)$ depends on x, which is the realization of the process $X_z(u)$ at z. No other relation between V and Z is present.

In this respect it is interesting to consider point-identification results from the literature on nonseparable triangular models, as Imbens & Newey (2009) or Torgovitsky (2015). In fact, every identification result in this literature in one way or another requires monotonicity of $h(x,\cdot)$ and $g(z,\cdot)$ in v and u, respectively (e.g. Imbens 2007 and references therein). The idea is that monotonicity makes h and g injective in v and u, respectively. In our setting, injectivity of h in v means that the paths of $Y_x(v)$ never intersect, so that for each (y,x) there is a unique v—this would guarantee that we could point-identify (under some more regularity assumptions) the latent distributions $P_{Y|X^*}$ and $P_{X|Z=z}$, as we would have a unique response profile (Y_x, X_z) for the observable $P_{Y,X|Z=z}$. This is why monotonicity is such a staple in the literature on point-identification of nonparametric models. △

Let us now show how to construct the indexing of the paths of the stochastic processes, adapting the ideas from Hess (1982). For this we first need to construct dyadic quasi-intervals in $\mathbb{R}^{[0,1]}$.

Lemma 2 (Construction of dyadic quasi-intervals)**.** *There is a family $(F_i)_{i\in\mathbb{N}^*}$ of dyadic quasi intervals F_i in $\mathbb{R}^{[0,1]}$ with the following properties for all $k\in\mathbb{N}$ and $(n_1,\dots,n_k)\in\mathbb{N}^k$.*

(i) $F_{n_1,\dots,n_k}$ is a dyadic quasi-interval of order $k-1$.

(ii) $\mathbb{R}^{[0,1]}=\bigcup_{n=1}^{\infty}F_n$. Moreover, the F_i can be made mutually disjoint via

$$D_{n_1,\dots,n_k} \coloneqq F_{n_1,\dots,n_k}\setminus(F_{n_1,\dots,n_k-1,1}\cup\dots\cup F_{n_1,\dots,n_{k-1},n_k-1}),$$

so that $D_{n_1,\dots,n_k}=\bigcup_{n=1}^{\infty}F_{n_1,\dots,n_k,n}$, $\mathbb{R}^{[0,1]}=\bigcupdot_n D_n$, and $P(F_{n_1,\dots,n_k})=P(D_{n_1,\dots,n_k})$.[15]

(iii) $(F_{n_1,\dots,n_k})^\circ=\emptyset$.

Proof. The proof in Hess (1982, p. 338–340) is for the space $\mathbb{R}^{(0,1]}\times\{0\}$, but the same construction works for $\mathbb{R}^{[0,1]}$ and therefore $[0,1]^{[0,1]}$. □

Dyadic quasi intervals form a countable partition of $\mathbb{R}^{[0,1]}$ and are the key step in constructing the indexing. They make the space $\mathbb{R}^{[0,1]}$ manageable by dividing it up into an ever finer partition of quasi-intervals, depending on how far one pushes the dyadic expansion. In particular, for each $k\in\mathbb{N}$ one obtains an infinite set of quasi-intervals at the k^{th} dyadic expansion. For $k=3$, for instance, one would have nine partitions of $\mathbb{R}$ into an infinite sequence of dyadic intervals[16] at the points $t_j=\frac{j}{2^m}$, $j=0,1,2,\dots,8$. The maximal length of each of those dyadic intervals in $\mathbb{R}$ would be less than or equal to $\frac{1}{8}$. The higher k, the finer the partition of $\mathbb{R}$ as well as the partition of $[0,1]$ by t. As $k\to\infty$, the sequence of quasi-intervals grows dense in $\mathbb{R}^{[0,1]}$.

[15] $\bigcupdot_n$ defines disjoint unions.

[16] Here, it is crucial to note the distinction between dyadic intervals and dyadic *quasi*-intervals. The former form a partition of the space $\mathbb{R}$ and hence form *a part* of a dyadic quasi-inteval. In fact, dyadic quasi-intervals are a "collection" of dyadic partitions of $\mathbb{R}$ at dyadic points in $[0,1]$.

The next step in the construction of indices is to actually construct, for both processes $Y_x(v)$ and $X_z(u)$ respectively, a Borel homeomorphism from the irrational numbers $\mathbb{I}$ in $[0,1]$ to the set of Hölder continuous functions on $[0,1]$, i.e. $C^{0,\lambda_Y}([0,1])$ and $C^{0,\lambda_X}([0,1])$ for λ_Y, λ_X defined in Proposition 2. We can construct these homeomorphisms again in analogy to Hess (1982). Recall in this respect that the set of Hölder continuous functions on $[0,1]$ is an F_σ subset of $C([0,1])$.

Lemma 3 (Construction of the indexings)**.** *For λ_Y and λ_X there exists a respective increasing sequence of closed subsets $(J_m^Y)_{m\in\mathbb{N}}$ of $\mathbb{I}$ and a respective $(1,1)$-homeomorphism $\phi_V : J^Y \to C^{0,\lambda_Y}([0,1])$, $J^Y \coloneqq \bigcup_{m=1}^\infty J_m^Y$ as well as $\phi_U : J^X \to C^{0,\lambda_X}([0,1])$, $J^X \coloneqq \bigcup_{m=1}^\infty J_m^X$ such that, for every $m \in \mathbb{N}$, the restriction of ϕ_V to J_m^Y is a $(0,1)$-homeomorphism from J_m^Y onto $C_m^{0,\lambda_Y}([0,1])$ and the restriction of ϕ_U to J_m^X is a $(0,1)$-homeomorphism from J_m^X onto $C_m^{0,\lambda_X}([0,1])$.*[17]

Proof. See Hess (1982, p. 340–341). □

Having constructed the dyadic quasi-intervals with the help of Lemma 2, the idea of Lemma 3 is to construct the two indexings, i.e. $(1,1)$-homeomorphisms by considering all non-empty intersections of the closed subsets $C_m^{0,\lambda}([0,1])$ of λ-Hölder continuous functions with the respective partition of quasi-intervals $F_{n_1,\dots,n_k}$. Formally, this can be expressed as

$$B^m_{n_1,\dots,n_k} \coloneqq F_{n_1,\dots,n_k} \cap C_m^{0,\lambda}([0,1]).$$

The respective indexing is then obtained by identifying the set of irrational numbers $\mathbb{I}$ with the Baire space $\mathbb{N}^\mathbb{N}$, the space of natural sequences (Aliprantis & Border 2006, Section 3.14) and to set

$$\mathbb{I} \supset J_m \coloneqq \{(n_i)_{i\in\mathbb{N}} \in \mathbb{N}^\mathbb{N} : B^m_{n_1,\dots,n_k} \neq \emptyset\},$$

which identifies the dyadic intervals through which Hölder continuous paths go with a subspace of Baire space and hence of the irrational numbers in $\mathbb{I}$. The proof of Lemma 3 in Hess (1982) then shows that this produces a $(1,1)$-homeomorphism from $J = \bigcup_{m=1}^\infty J_m$ onto $C^{0,\lambda}([0,1])$. Note that $C^{0,\lambda}([0,1]) \subset C([0,1])$.

When facing binary or discrete Z one can adjust the above construction rather easily and define X_z to be a "multi-point-process" in the discrete case, by choosing only d points on the unit interval for Z and not a complete dyadic approximation. This would make the construction much easier in fact; all one has to do is make sure that the stochastic process X_z is well-defined in a mathematical sense in order to define a joint measure μ over both processes. We will not go into detail here since our focus is on the most general (and most complicated) case where all variables are allowed to be continuous.

Theorem 1 allows us to obtain general causal effects if we apply it at two different points x_0 and x_1 for the same y^* by taking the difference $F_{Y|X^*=x_1}(y^*) - F_{Y|X^*=x_0}(y^*)$ as the objective

[17] $C_m^\lambda([0,1])$ is the set of all functions f on $[0,1]$ for which $|f(x_1) - f(x_2)| \leq m|x_1 - x_2|^\lambda$ for all $x_1, x_2 \in [0,1]$. See the notation in the appendix for further information.

function. The value of this problem then is the distributional causal effect when X is changed from x_0 to x_1. Similarly, one could obtain bounds on causal effects for subsets of X by integrating over some Borel sets E_{x_0} and E_{x_1}, for instance, in which case the causal effect would be obtained by using the objective function

$$\int_{[0,1]} \int_{E_{x_0}} H_{Y,X}(y^*, x, v) dx \pi_1 \mu(dv) - \int_{[0,1]} \int_{E_{x_1}} H_{Y,X}(y^*, x, v) dx \pi_1 \mu(dv).$$

Theorem 1 is hence in the form to obtain any causal effect. For the ATE we also have to adjust the objective function a little.

Corollary 1 (Bounds on the ATE)**.** *In order to obtain upper bounds on the ATE of a change from $X = x_0$ to $X = x_1$ one needs to change the objective function to*

$$\int_{[0,1]} H^*(x_1, v) \pi_1 \mu(dv) - \int_{[0,1]} H^*(x_0, v) \pi_1 \mu(dv),$$

for $H^*(x, v) \coloneqq Y_x(v)$.

Throughout the next section we focus on the programs (3) where the objective function is for obtaining probabilities and not the ATE. Since the objective function for the ATE is continuous all results we prove in the following for the distributional causal effects carry over to the ATE.

Before turning to properties of the programs, we want to state an equivalence result. Note that we have stated the programs (3) in such a way that they replicate the observable distribution function $F_{Y,X|Z=z}$, which is a more general requirement than the framework in Balke & Pearl (1994) or Kitagawa (2009) who work with the density function $f_{Y,X|Z=z}$. The following proposition shows that we can obtain the bounds by considering either the CDF or the PDF if the latter exists.

Proposition 3 (Equivalence result for densities)**.** *If $F_{Y,X|Z=z}$ possesses a density $f_{Y,X|Z=z}(y, x)$ with respect to Lebesgue measure for almost every $z \in [0, 1]$, then the following programs are almost everywhere equivalent to the respective programs* (3)*:*

$$\begin{aligned} &min/max_{\mu \in \mathcal{P}^*([0,1]^2)} \quad \int_{[0,1]} H_{Y,X}(y^*, x_0, v) \pi_1 \mu(dv) \\ &s.t.\ f_{Y,X|Z=z}(y, x) = \int \int \Gamma(y, x, z, v, u) \mu(dv, du), \end{aligned} \tag{7}$$

where

$$\Gamma(y, x, z, v, u) = \delta\left(y - Y_{X_z(u)}(v)\right) \delta\left(x - X_z(u)\right)$$

is the product of two shifted Dirac delta-distributions.

When establishing the theoretical properties of the optimization programs in the following we use the CDF-versions, because it will be easier to establish the large sample distributions of the

programs when estimating $F_{Y,X|Z=z}$ by a smoothed estimator $\hat{F}_{Y,X|Z=z;h_n}$ with some bandwidth h_n; the additional smoothness makes proving the large sample results easier for us in section 4.4.

Let us first turn to analyzing the infinite dimensional linear programs in more depth.

4.3 Theoretical properties of the linear programs

The maximal and minimal solutions to the infinite dimensional linear program established in the previous subsection provide the required bounds on $P_{Y|X^*=x}$ and *a fortiori* on the ATE. We therefore need to analyze its properties to guarantee that the optimization problems are well-behaved. We do so in this and the next subsection. In this subsection we focus on mathematical, in the next subsection on statistical properties.

We show that the constraint set, which in actuality is a constraint correspondence[18] as it depends on the observed distribution $F_{Y,X|Z}$, is convex and pre-compact-valued in the weak topology. Furthermore, we establish the dual problems and show that there is no duality gap under a weak structural assumption on the estimable $F_{Y,X|Z}$, so that one can use the dual program to solve the problem. Most importantly, however, we show that the optimization problems in Theorem 1 take the form of a general version of the *general capacity problem* (Anderson & Nash 1987; Anderson, Lewis & Wu 1989), a well-studied infinite-dimensional linear program in measure spaces which takes the form

$$\begin{aligned} &\min_{\mu\in\mathcal{M}(\mathcal{Y})} \int_{\mathcal{Y}} f(y)\mu(dy) \\ &\text{s.t.} \quad \int_{\mathcal{Y}} \varphi(y,x)\mu(dy) \geq g(x), \end{aligned} \tag{8}$$

where it is usually assumed that $f \in C(\mathcal{Y})$, $g \in C(\mathcal{X})$.[19]

The general capacity problem extends the simple capacity problem from Choquet (1954). Our problem is even more general than the general capacity problem as we integrate over paths of stochastic processes and not just points in Euclidean space. Therefore, we solve this problem approximately in practice. We do this in section 5 by a sampling approach; in addition, we borrow the idea of discarding unimportant paths for our optimization from cutting plane approaches designed for general capacity problems on Euclidean spaces, introduced in generality by Lai & Wu (1992) and Wu, Fang & Lin (2001).

Let us start with phrasing the problems (3) in terms of the general capacity problem. Comparing (3) and (8) we see that our problem is already in the form of the general capacity problem, except for the fact that we have an equality constraint and require $\mu \in \mathcal{P}^*([0,1]^2)$ instead of $\mathcal{M}([0,1]^2)$. Moreover, the functions $H_{Y,X}$ and $G_{Y,X}$ do not lie in $C([0,1]^2)$ and $C([0,1]^5)$, respectively, as they are (products of) indicator functions. We can, however, approximate $H_{Y,X}$ and

[18] A correspondence is a "multivalued function". For a nice introduction to correspondences, consider Aliprantis & Border (2006, Chapter 17).

[19] Note that if we work in more general spaces like $D([0,1])$ or $\mathbb{R}^{[0,1]}$, this assumption is unrealistic. One then has to prove analogous results to the ones we derive below for $F_{Y,X|Z=z}(y,x) \in D([0,1]^3)$ for instance, which proceeds along the same lines if we equip $D([0,1]^3)$ with the Skorokhod metric.

$G_{Y,X}$ by (products of) smooth functions such as versions of the logistic function $S(x) = \frac{\exp(x)}{\exp(x)+1}$. For example, we can approximate $\mathbb{1}_{[0,y]}\{Y_x(v)\}$ by

$$S_1(Y_x(v), y, \varepsilon_1) \coloneqq \frac{1}{\left(1 + \exp\left(-\varepsilon_1\left(Y_x(v) + \varepsilon_1^{-1/2}\right)\right)\right)\left(1 + \exp\left(-\varepsilon_1\left(y - Y_x(v) + \varepsilon_1^{-1/2}\right)\right)\right)}$$

and $\mathbb{1}_{[0,y]\times[0,x]}\{Y_{X_z(u)}(v), X_z(u)\}$ by $S_1(Y_{X_z(u)}(v), y, \varepsilon_1) \cdot S_2(X_z(u), x, \varepsilon_2)$ for $\varepsilon_1, \varepsilon_2 > 0$ and as $\varepsilon_1, \varepsilon_2 \to +\infty$.

Also recall that $Y_x(v)$ and $X_z(u)$ are $(0,1)$-homeomorphisms between $[0,1]$ and $C_m^{0,\lambda}([0,1])$ for every $m \in \mathbb{R}^+$ by the construction in Lemmas 2 and 3, so they are continuous maps in particular. The compositions $S_1(Y_x(v), \varepsilon_1)$ and $S_1(Y_x(v), \varepsilon_1) \cdot S_2(X_z(u), \varepsilon_2)$ are hence continuous in u and v for all $\varepsilon_1, \varepsilon_2 > 0$. From now on we will denote the continuous approximation of $H_{Y,X}(y, x, v)$ by $\Xi(y, x, v, u) \coloneqq S_1(Y_x(v), \varepsilon_1)$ and the continuous approximation of $G_{Y,X}(y, x, z, v, u)$ by $\Theta(y, x, z, v, u) \coloneqq S_1(Y_x(v), y) \cdot S_2(X_z(u), x)$, suppressing the dependence on ε_1 and ε_2 and letting Ξ depend on u. Analogously, we can approximate the Dirac delta-functions $\Gamma(y, x, z, v, u)$ by standard mollifiers, i.e. kernel density estimators in order to translate our problems (7) to the setting of the general capacity problem. Again, the density case is more natural for the general capacity problem, but we focus on the slightly more general case of the CDF since our practical problem in section 5 is based on the CDF version.

The fact that the kernels $H_{Y,X}$ and $G_{Y,X}$ can be approximated by continuous functions helps us introduce appropriate operators. In particular, we can define a linear integral operator $\Theta\mu$ by

$$\Theta\mu(y, x, z) \coloneqq \int_{[0,1]^2} \Theta(y, x, z, v, u)\mu(dv, du).$$

The domain of Θ is $\mathcal{M}([0,1]^2)$ and the codomain is $C([0,1]^3)$ since Θ is the continuous approximation of an indicator function. In this setting we can define the bilinear functionals $\langle \rho, f\rangle_1$ on $\mathcal{M}([0,1]^2) \times C([0,1]^2)$ and $\langle \widetilde{f}, \nu\rangle_2$ on $C([0,1]^3) \times \mathcal{M}([0,1]^3)$ by

$$\begin{aligned}\langle \rho, f\rangle_1 &\coloneqq \int_{[0,1]^2} f(v, u)\rho(dv, du) \quad \text{and} \\ \langle \widetilde{f}, \nu\rangle_2 &\coloneqq \int_{[0,1]^3} \widetilde{f}(y, x, z)\nu(y, x, z),\end{aligned}$$

so that we have by the Fubini-Tonelli theorem

$$\begin{aligned}\langle \Theta\mu, \nu\rangle_2 &= \int_{[0,1]^3}\int_{[0,1]^2} \Theta(y, x, z, v, u)\mu(dv, du)\nu(dy, dx, dz) \\ &= \int_{[0,1]^2}\int_{[0,1]^3} \Theta(y, x, z, v, u)\nu(dy, dx, dz)\mu(dv, du) = \langle \mu, \Theta^*\nu\rangle_1,\end{aligned}$$

where $\Theta^*\nu : \mathcal{M}([0,1]^3) \to C([0,1]^2)$ is the adjoint operator of $\Theta\mu$. We always endow $C([0,1]^3)$

with the uniform topology, but will consider different topologies on $\mathcal{M}([0,1]^2)$. From now on, we will prove all properties of the following alternative versions of (3)

$$\begin{aligned}
&\inf_{\mu\in\mathcal{P}^*([0,1]^2)} \quad \int_{[0,1]} \Xi(y^*, x_0, v, u)\mu(dv, du) \\
&\text{s.t. } F_{Y,X|Z=z}(y,x) = \int_{[0,1]^2} \Theta(y,x,z,v,u)\mu(dv,du) \qquad \text{and} \\
&\sup_{\mu\in\mathcal{P}^*([0,1]^2)} \quad \int_{[0,1]} \Xi(y^*, x_0, v, u)\mu(dv, du) \\
&\text{s.t. } F_{Y,X|Z=z}(y,x) = \int_{[0,1]^2} \Theta(y,x,z,v,u)\mu(dv,du).
\end{aligned} \tag{9}$$

As a first step towards showing that the problems (9) are well-defined, we need to prove that the constraint correspondence[20]

$$\begin{aligned}
\mathcal{A}(F_{Y,X|Z=z}(y,x)) &\coloneqq \mathcal{P}^*([0,1]^2) \cap \mathcal{E}([0,1]^2) \\
&= \mathcal{P}^*([0,1]^2) \cap \left\{\mu \in \mathcal{P}([0,1]^2) : \Theta\mu = F_{Y,X|Z=z}(y,x)\right\}
\end{aligned} \tag{10}$$

is non-empty.

Lemma 4 (Non-emptiness of the constraint correspondence)**.** *If Assumptions 1 – 4 hold then there exist c_x, c_y, α, β, γ, $\delta > 0$ as required in Assumption 3 and a $\mu \in \mathcal{P}^*([0,1]^2)$ such that $\Theta\mu = F_{Y,X|Z=z}(y,x)$.*

From now on we will implicitly assume that the constants have been chosen such that the problems have a solution, as Assumptions 2 and 4 are theoretical devices for our proofs, but do not appear directly when solving the problem in practice. Based on the above results, we have the following

Proposition 4 (Regularity of the constraint correspondence)**.** *Under Assumptions 1 – 3 the set $\mathcal{P}^*([0,1]^2)$ is convex and compact in the weak topology. Under Assumptions 1 – 4 the constraint correspondence $\mathcal{A}(F_{Y,X|Z=z}(y,x))$ is non-empty, convex, and pre-compact for an appropriate choice of $\alpha, \beta, \gamma, \delta, c_y, c_x > 0$.*

Proposition 4 shows that the correspondence $\mathcal{A}(F_{Y,X|Z=z}(y,x))$ is well-behaved; in particular, the set $\mathcal{P}^*([0,1]^2)$ is convex and compact in the weak topology, even though its interior is empty. Convexity and compactness are very helpful properties, as they enable us to prove statistical large sample properties of the value functions $\underline{m}(F_{Y,X|Z=z}(y,x))$ and $\overline{m}(F_{Y,X|Z=z}(y,x))$, which we define as

$$\underline{m}(F_{Y,X|Z=z}(y,x)) = \begin{cases} \inf_\mu \Xi\mu & \text{if} \quad \mu \in \mathcal{A}(F_{Y,X|Z=z}(y,x)) \\ +\infty & \text{otherwise} \end{cases}$$

[20] $\mathcal{A}(F_{Y,X|Z=z}(y,x))$ is indeed a correspondence as it depends on $F_{Y,X|Z=z}(y,x)$.

$$\overline{m}(F_{Y,X|Z=z}(y,x)) = \begin{cases} \sup_\mu \Xi\mu & \text{if} \quad \mu \in \mathcal{A}(F_{Y,X|Z=z}(y,x)) \\ -\infty & \text{otherwise} \end{cases}.$$

Before turning to this we provide the dual problems to (9) and show that they yield the same results as the primal problems, for the sake of completeness. In finite dimensions it does not matter if one solves the primal problem or its dual, since both will give the same solution under fairly weak and standard conditions; this property is called strong duality. In contrast to the finite dimensional case, strong duality need not hold in infinite dimensions. We now show, however, that under the above assumptions strong duality does hold for the problems (9).

Proposition 5 (Lagrangian dual programs and strong duality)**.** *The Lagrangian dual problems to* (9) *are*

$$\begin{aligned} &\sup_{\nu\in\mathcal{M}([0,1]^3)} \inf_{\mu\in\mathcal{P}^*([0,1]^2)} \int_{[0,1]^2} \Xi(y^*, x_0, v, u)\mu(dv, du) \\ &\quad + \int_{[0,1]^3} \left(F_{Y,X|Z=z}(y,x) - \int_{[0,1]^2} \Theta(y,x,z,v,u)\mu(dv,du) \right) \nu(dy,dx,dz) \\ &\textit{and} \qquad\qquad (11) \\ &\inf_{\nu\in\mathcal{M}([0,1]^3)} \sup_{\mu\in\mathcal{P}^*([0,1]^2)} \int_{[0,1]^2} \Xi(y^*, x_0, v, u)\mu(dv, du) \\ &\quad + \int_{[0,1]^3} \left(F_{Y,X|Z=z}(y,x) - \int_{[0,1]^2} \Theta(y,x,z,v,u)\mu(dv,du) \right) \nu(dy,dx,dz). \end{aligned}$$

Also, there is no duality gap under Assumptions 1 – 4.

We do not need the dual program for our purposes, but the proof of strong duality in our setting is instructive—in fact, we can build a proof for the large sample distribution of the infinite dimensional programs based on ideas of the proof of strong duality. Also note that Proposition 5 does not assert that the optimal values of the dual problems are achieved by some $\nu \in \mathcal{M}([0,1]^3)$. The issue is that we can only prove that the optimal value functions $\overline{m}(F_{Y,X|Z=z}(y,x))$ and $\underline{m}(F_{Y,X|Z=z}(y,x))$ are upper- respectively lower semicontinuous, which guarantees strong duality but not the fact that the optima of the dual problems are actually achieved. This is not an issue for us since we only care about the value functions in our setting.

The problem which prevents $\overline{m}(F_{Y,X|Z=z}(y,x))$ and $\underline{m}(F_{Y,X|Z=z}(y,x))$ from being continuous in our setting is the constraint that $\mu \in \mathcal{P}^*([0,1]^2)$, which is a subset with empty interior in $\mathcal{M}([0,1]^2)$ equipped with either the total variation distance or the weak topology. This means that for any conditional distribution function $F'_{Y,X|Z=z} \in C([0,1]^3)$ which does not correspond to a probability measure but a general Borel measure, we have $\overline{m}(F'_{Y,X|Z=z}(y,x)) = -\infty$ and $\underline{m}(F'_{Y,X|Z=z}(y,x)) = +\infty$ as $\mathcal{A}(F'_{Y,X|Z=z})$ is empty in this case. The set of all continuous conditional probability distribution functions $F_{Y,X|Z=z}$ also has empty interior in the set of all

continuous functions, so that the interior of the domain of $\overline{m}$ and $\underline{m}$ is empty, which prevents them from being continuous on all of $C([0,1]^3)$. We therefore cannot simply use the infinite dimensional Slater condition which would immediately give strong duality. However, continuity of $\underline{m}$ and $\overline{m}$ is reestablished if we restrict $F_{Y,X|Z}$ to satisfy Assumption 4, which we denote by $F_{Y,X|Z} \in \mathcal{F}([0,1]^3) \subset C([0,1]^3)$ for $\mathcal{F}([0,1]^3) \coloneqq \{F \in C([0,1]^3) : F \text{ satisfies Assumption 4}\}$, and this is the idea for the proof of asymptotic normality, which we turn to now.

4.4 Statistical properties of the linear programs

Let us now turn to the statistical properties of the programs. In the following we prove that the natural plug-in estimators where we replace $F_{Y,X|Z=z}$ by a smoothed estimator $\hat{F}_{Y,X|Z=z;h_n} \in \mathcal{F}([0,1]^3)$ are well-behaved in the sense that they are uniformly asymptotically linear. We explicitly require $\hat{F}_{Y,X|Z=z;h_n}$ to be continuous and to satisfy Assumption 4. In addition, for this section we assume that $P_{Y,X|Z=z}$ possesses a density with respect to Lebesgue measure for almost every $z \in \mathcal{Z}$, which allows us to use standard results from the literature of smoothed empirical processes, in particular from Giné & Nickl (2008).

Assumption 5. *The observable measure $P_{Y,X|Z=z}$ possesses a density $p_{Y,X|Z=z}$ with respect to Lebesgue measure for almost every $z \in \mathcal{Z}$.*

We denote our kernel density estimator for $p_{Y,X,Z}$ by

$$\hat{P}_{Y,X,Z;n} * K_{h_n}(y,x,z) \coloneqq \frac{1}{nh_n^3} \sum_{i=1}^{n} K\left(\frac{y-Y_i}{h_n}\right) K\left(\frac{x-X_i}{h_n}\right) K\left(\frac{z-Z_i}{h_n}\right),$$

where K denotes a smoothing kernel, h_n denotes the bandwidth, and $f * g$ denotes the convolution between two functions f and g. Based on this, we have the following notation for the conditional kernel density estimator for $p_{Y,X|Z}$

$$\hat{P}_{Y,X|Z;n} * K_{h_n}(y,x,z) \coloneqq \frac{\frac{1}{nh_n^3} \sum_{i=1}^{n} K\left(\frac{y-Y_i}{h_n}\right) K\left(\frac{x-X_i}{h_n}\right) K\left(\frac{z-Z_i}{h_n}\right)}{\frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{z-Z_i}{h_n}\right)}.$$

Using this notation we write the smoothed empirical process as

$$\sqrt{n}(\hat{P}_{Y,X|Z;n} * K_{h_n} - P_{Y,X|Z})(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(f * K_{h_n}(Y_i, X_i, Z_i) - \int_{[0,1]^3} f(y,x,z) P_{Y,X|Z=z}(dy, dx) \right)$$

for functions f in some appropriate (Donsker or pregaussian) class $\mathcal{G}$. Weak convergence of this empirical process to some limiting process $\mathbb{Z}$ over the class $\mathcal{G}$ is then denoted as

$$\sqrt{n}\left(\hat{P}_{Y,X|Z;n} * K_{h_n} - P_{Y,X|Z}\right) \overset{\ell^\infty(\mathcal{G})}{\Longrightarrow} \mathbb{Z},$$

where $\ell^\infty(\mathcal{G})$ denotes the space of all uniformly real bounded functions on $\mathcal{G}$. If we set the class $\mathcal{G}$ to be the set of all rectangles in $[0,1]^3$, then the smoothed empirical process above is a smoothed empirical distribution function, and we write

$$\sqrt{n}\left(\hat{F}_{Y,X|Z;h_n} - F_{Y,X|Z}\right) \Longrightarrow \mathbb{Z},$$

where the limiting process $\mathbb{Z}$ is indexed by all rectangles on $[0,1]^3$.

In order to make the smoothed empirical processes asymptotically linear, we require slightly stronger assumptions on the distribution function than we had previously.

Assumption 6 (Regularity of the density)**.** *The density $p_{Y,X|Z}$ associated with $P_{Y,X|Z}$ is bounded on $[0,1]^2$ for almost every $z \in [0,1]$ or satisfies $p_{Y,X|Z} \in C^{s,\lambda'}([0,1]^3)$ for some $s \geq 0$ and $\lambda' \geq 0$.*

We also choose appropriate kernels K_{h_n} and bandwidths h_n.

Assumption 7 (Regularity of the kernel density estimator)**.** *The kernels K_{h_n} are of order $r = \lambda' + 1 - t$ for some t with $0 < t < \lambda' + 1$. The bandwidths $h_n > 0$ satisfy $h_n^{\lambda'+1-t} n^{1/2} \to 0$ as $n \to \infty$.*

Assumptions 6 and 7 guarantee that the smoothed empirical processes converge, which follows from Proposition 4 in Giné & Nickl (2008); in particular, the choice of bandwidth and the kernel make the bias asymptotically negligible. These are all standard assumptions and very weak.

Finally, we assume that $\hat{F}_{Y,X|Z;h_n} \in \mathcal{F}([0,1]^3)$, so that the optimization problems are guaranteed to be non-empty.

Assumption 8 (Continuous path requirement for the smoothed empirical distribution)**.** $\hat{F}_{Y,X|Z;h_n} \in \mathcal{F}([0,1]^3)$*, i.e.*

$$\int_{[0,1]^2}\int_{[0,1]^2} |(s_y,s_x)-(t_y,t_x)|^{\eta_1} d\hat{P}_{Y,X|Z=z_1;n} * K_{h_n}(s_y,s_x) d\hat{P}_{Y,X|Z=z_2;n} * K_{h_n}(t_y,t_x) \leq c_{y,x}|z_1 - z_2|^{1+\eta_2}$$

for some $c_{y,x}, \eta_1, \eta_2 > 0$.

Assumption 8 is a sufficient condition for the linear programs to have a solution in finite samples. In particular, it guarantees that the value functions are Hadamard differentiable tangentially to $\mathcal{F}([0,1]^3)$. Assumption 8 is simply a regularity condition and a very weak one at that. It should be satisfied in nearly all practical settings; in those where it is not one should work with more general stochastic processes that allow for jumps. We now have the following

Proposition 6 (Uniform large sample results)**.** *Under Assumptions 1 – 8 it holds that*

$$\sqrt{n}\left(\overline{m}(\hat{F}_{Y,X|Z;h_n}) - \overline{m}(F_{Y,X|Z})\right) \Longrightarrow \dot{\overline{m}}_{F_{Y,X|Z}}(\mathbb{G}) \quad \textit{and}$$
$$\sqrt{n}\left(\underline{m}(\hat{F}_{Y,X|Z;h_n}) - \underline{m}(F_{Y,X|Z})\right) \Longrightarrow \dot{\underline{m}}_{F_{Y,X|Z}}(\mathbb{G})$$

where $\mathbb{G}$ *a Brownian bridge indexed by all rectangles on* $[0,1]^3$*, and*

$$\dot{\overline{m}}_{F_{Y,X|Z}}(F) = \min_{\nu \in \partial \overline{m}(F_{Y,X|Z})} \langle F, \nu \rangle_2 \qquad \textit{for} \quad F \in \mathcal{F}([0,1]^3)$$

$$\dot{\underline{m}}_{F_{Y,X|Z}}(F) = \max_{\nu \in \partial \underline{m}(F_{Y,X|Z})} \langle F, \nu \rangle_2 \qquad \textit{for} \quad F \in \mathcal{F}([0,1]^3)$$

are the directional Hadamard derivatives of $\overline{m}(\cdot)$ *and* $\underline{m}(\cdot)$ *at* $F_{Y,X|Z}$ *tangentially to the set* $\mathcal{F}([0,1]^3)$*.*

Here $\partial \underline{m}(F_{Y,X|Z})$ denotes the subgradient of $\underline{m}$ at $F_{Y,X|Z}$. In general, the measure $\nu \in \mathcal{M}([0,1]^3)$ belongs to $\partial \underline{m}(F_{Y,X|Z})$ if $\underline{m}(F_{Y,X|Z} + tF) \geq \underline{m}(F_{Y,X|Z}) + t\langle F, \nu \rangle_2$ for all $F \in C([0,1]^3)$ and $t > 0$, see for instance Rockafellar (1974, p. 33). Note furthermore that $\max_{\nu \in \partial \underline{m}(F_{Y,X|Z})} \langle F, \nu \rangle_2$ is the support function of the subgradient, which is equicontinuous in a neighborhood around zero in our setting (see p. 31 in Rockafellar 1974). Also note that the value functions $\underline{m}$ and $\overline{m}$ are barely lower- respectively upper semicontinuous on $C([0,1]^3)$ and clearly not Hadamard differentiable there. We only require differentiability *tangentially to* $\mathcal{F}([0,1]^3) \subset C([0,1]^3)$, however, which is enough for the Functional Delta Method to hold, which in turn proves asymptotic linearity (Shapiro 1991, Theorem 2.1).

To give an intuitive explanation of this result, note that we can only show that the linear programs perturb continuously over the set $\mathcal{F}([0,1]^3)$, but not over all of $C([0,1]^3)$ as there are some directions in which the optimal value of the programs change drastically even though $\hat{F}_{Y,X|Z=z;h_n}$ only changes slightly. We therefore provide directions, namely directions in $\mathcal{F}([0,1]^3)$, for which the optimal value does vary continuously, which we do by showing that the value function is both upper- and lower semicontinuous on this set. This is where Assumption 8 comes into play, which requires that there always exists a solution to the finite sample linear programs. The functional Delta-Method is still valid if we only consider certain directions and not the whole space, so that we can use it to derive the large sample properties in particular directions.

Proposition 6 can be used to perform inference on each bound separately. In practice, the large sample distribution will be estimated by bootstrapping or subsampling methods. It is known, however, that the bootstrap for the delta method can fail in general when the function is only directionally Hadamard differentiable. Dümbgen (1993) and more recently Fang & Santos (2014) derive results which provide consistency of (versions of) the bootstrap in this setting. Second, note that Proposition 6 gives large sample results for each bound separately. In order to obtain uniform confidence intervals with appropriate coverage, one can use established ideas from the literature (Chernozhukov, Lee & Rosen 2013, Imbens & Manski 2004, Stoye 2009). The results developed in this section fit together nicely with the existing literature on uniform inference in partially identified models, and we refer to these articles for further information, as this is not the main focus of this article.

Let us finally turn to the practical implementation and the application.

5 Practical implementation

We want to apply the results from the previous section to nonparametrically estimate bounds on distributional causal effects for expenditures. For this we have to introduce the practical implementation first, which we do in this section.

The infinite dimensional linear programs we constructed theoretically in the previous section are similar to, but more general than, optimal stochastic control or reinforcement learning problems since we do not assume a Markovian structure on the stochastic processes in general. Finding the optimal global solution to these general problems via a brute force method requires exponential time and memory (e.g. Kappen 2007). Moreover, even checking whether a given solution to the problem is smaller than some value requires us to check the solution on exponentially many paths, which is infeasible. In practice, we therefore solve them by *approximating* their solutions via a *sampling* approach, which we find yields very reasonable approximate solutions, already for coarse grids.

We now present an algorithm for solving the infinite dimensional linear programs in practice via this "sampling of paths" approach. Recall from the previous section that we have to perform a dyadic decomposition of the unit interval, and the finer the dyadic decomposition, the more accurate the solution to the optimization program. Throughout this section we denote by $m_i = 2^i + 1$ the total number of points in the respective dyadic decomposition of order $i \in \mathbb{N}$. As $i \to \infty$, we obtain the true infinite dimensional linear programs from section 4. Of course, there is a tradeoff in practice as a finer dyadic decomposition of the unit interval leads to a higher complexity of the linear programs to solve.

In practice, there are many different ways to sample paths of stochastic processes. Standard ways are to use general stochastic differential equations driven by Levy-processes for instance. Another approach—specifically for sampling continuous paths—is to use a recombining trinomial tree, which is centered on the m_i points of the respective dyadic decomposition. Other assumptions besides continuity would translate to different requirements on the paths of the stochastic processes. For instance, more stringent assumptions like monotonicity can be incorporated by ruling out non-monotonic paths in this model. This is one way of changing the set $\mathcal{P}^*([0,1]^2)$ from Theorem 1 in practice.

Note, however, that the way one samples the paths of the processes has an effect on the solution, which should be taken into account when applying the algorithm in practice. In particular, one can introduce assumptions on the paths of the processes by choosing different sampling approaches. For instance, using an SDE driven by Brownian motion for the sampling will only allow for stochastic paths of this structure and not other approaches. It is therefore important in practice to be clear which approach one uses for the sampling.

In this paper we use trinomial trees for sampling paths of stochastic processes as we want to uphold continuity as an assumption. To give an example of our approach consider the simple example by setting $i = 2$ in the dyadic approximation, which gives $m_2 = 5$ points, i.e. the points

$0, 0.25, 0.5, 0.75, 1$. We decompose all intervals for Y, X, and Z like this. Consider the recombining tree for the stochastic process Y_x. At the point $x = 0$ Y consequently has 5 possible starting points, $y \in \{0, 0.25, 0.5, 0.75, 1\}$. Suppose we pick $y_0 = 0.5$ at $x = 0$. Then we construct the trinomial tree by allowing y to take three possible values at the point $x = 0.25$: it can stay at its place, in which case, $y_{0.25} = 0.5$, it can move up one node, i.e. $y_{0.25} = 0.75$, or it can move down one node, i.e. $y_{0.25} = 0.25$. If we move through the tree for all points like this, we have sampled one path of the respective continuous stochastic process at those points. The same holds for the process X_z.

When sampling paths from this trinomial tree we use a general "sampling" probability which we denote by P_s. In particular, we randomly determine the probability for each of the three possible paths (up one node, stay the same, down one node) at each node anew; that is, P_s is defined as the process which gives, at each node, a new randomly assigned probability distribution over the three alternatives. We do this for all samples of nodes and think that this gives the greatest variety of paths in our sampling approach, so that our only assumption really just is continuity and nothing else. In principle, we could use other sampling probabilities P_s, but these would most likely translate to further, in our case unwanted, assumptions. For instance, a fixed probability distribution which put the most weight on the "up" alternative would put most weight on increasing paths; this would result in most paths being increasing as the probability of sampling a decreasing path in practice would be very low.

We can now present our algorithm for solving the infinite dimensional linear programs approximately in practice.

Algorithm.

0. Initial step: randomly sample some set

$$\mathcal{R}_0 \coloneqq \{(Y_{x_i}(v_l), X_{z_i}(u_l)), l = 1, \ldots, k_{init}, i = 1, \ldots, m_j\}$$

of initial paths, where k_{init} is the number of initial paths to sample, and where m_j is the number of grid-points on the unit interval based on the dyadic decomposition of order j. Sample paths with or without replacement[21] *with sampling measure P_s. Fix some $\delta > 0$ and $n_\delta \in \mathbb{N}$, which will control the convergence criterion of the algorithm. Compute the matrix Θ_0 and the corresponding vector Ξ_0, based on these paths, using the logistic approximations $S_1(Y_x(v), y, \varepsilon_1)$ and $S_1(Y_{X_z(u)}(v), y, \varepsilon_1) \cdot S_2(X_z(u), x, \varepsilon_2)$ for some large $\varepsilon_1, \varepsilon_2 > 0$. Set $k = 1$ and set*

$$\Theta_0^{max} = \Theta_0^{min} = \Theta_0 \qquad \text{and} \qquad \Xi_0^{max} = \Xi_0^{min} = \Xi_0.$$

[21] In our application we sample with replacement, because we want to test whether this algorithm converges even if we are able to sample all possible paths. We are in the process of implementing a way to sample without replacement, which is most likely more efficient.

1. *At iteration k, randomly sample a set*

$$\mathcal{R}_k \coloneqq \left\{ \left(Y_{x_j}(v_l), X_{z_j}(u_l) \right), l = 1, \ldots, k_{add}, i = 1, \ldots, m_j \right\}$$

of stochastic paths to add to the program, where k_{add} is the number of paths to add. Sample paths with or without replacement under P_s and make sure the sampled paths are unique. Compute the preliminary matrices $\widetilde{\Theta}_k^{min}$ and $\widetilde{\Theta}_k^{max}$ as well as the vectors $\widetilde{\Xi}_k^{min}$ and $\widetilde{\Xi}_k^{max}$ based on these paths as in the initial step. Update the matrices Θ_{k-1}^{min} and Θ_{k-1}^{max} as

$$\Theta_k^{min} = [\Theta_{k-1}^{min} \;\; \widetilde{\Theta}_k^{min}], \qquad \Theta_k^{max} = [\Theta_{k-1}^{max} \;\; \widetilde{\Theta}_k^{max}],$$

i.e. by appending the respective columns of $\widetilde{\Theta}_k$ to $\Theta_{k-1}^{\cdot}$. In addition, update the vectors Ξ_{k-1}^{min} and Ξ_{k-1}^{max} as

$$\Xi_k^{min} = [\left(\Xi_{k-1}^{min}\right)' \;\; \left(\widetilde{\Xi}_k^{min}\right)']' \qquad \textit{and} \qquad \Xi_k^{max} = [\left(\Xi_{k-1}^{max}\right)' \;\; \left(\widetilde{\Xi}_k^{max}\right)']',$$

where A' denotes the transpose of the matrix A.

2. *Solve the programs*

$$\begin{array}{lll} \underset{\mu \geq 0,\, \vec{1}'\mu \leq 1}{\textit{minimize}} & \left(\Xi_k^{min}\right)' \mu & \textit{and} \\ s.t. \quad \Theta_k^{min}\mu = \hat{F}_{Y,X|Z=z;h_n} & & \end{array} \qquad \begin{array}{l} \underset{\mu \geq 0,\, \vec{1}'\mu \leq 1}{\textit{maximize}} \quad (\Xi_k^{max})' \mu \\ s.t. \quad \Theta_k^{max}\mu = \hat{F}_{Y,X|Z=z;h_n}, \end{array} \tag{12}$$

where μ is a vector with dimension equal to the number of sampled paths, $\hat{F}_{Y,X|Z=z;h_n}$ is the smoothed estimator of the CDF from section 4 supported on the m_j points of the dyadic approximation of the unit interval, and where $\vec{1}$ denotes the vector of the same dimension as μ containing all ones. Store the optimal solutions to these problems as $V_{k,min}$ and $V_{k,max}$ and the optimizers as $\mu_{k,min}$ and $\mu_{k,max}$. If the moving standard deviations

$$\left(\frac{1}{n-1} \sum_{j=1}^{n_\delta} \left(V_{k-j,min} - \bar{V}_{k,min} \right)^2 \right)^{1/2} \leq \delta \quad \textit{and} \quad \left(\frac{1}{n-1} \sum_{j=1}^{n_\delta} \left(V_{k-j,max} - \bar{V}_{k,max} \right)^2 \right)^{1/2} \leq \delta,$$

for the window length n_δ and the $\delta > 0$ chosen in stage 0, stop and output $V_{k,min}$ and $V_{k,max}$ as the solution. Here,

$$\bar{V}_{k,min} = \frac{1}{n} \sum_{j=1}^{n_\delta} V_{k-j,min} \qquad \textit{and} \qquad \bar{V}_{k,max} = \frac{1}{n} \sum_{j=1}^{n_\delta} V_{k-j,max}$$

are the moving averages.

3. *Delete all columns from Θ_k^{min} and all rows from Ξ_k^{min} for which the corresponding values of*

$\mu_{k,min}$ are zero. Analogously for Θ_k^{max}. Increment $k \to k+1$ and go to step 1.

The basic idea of the algorithm is as follows: in principle, we would like to solve the linear optimization problems in the algorithm by optimizing over all possible continuous paths of the processes Y_x and X_z. However, even for rather coarse dyadic approximations this requires an exorbitant amount of memory. Our idea therefore is to randomly sample paths and to hope that we sample "enough" of the relevant paths in order to obtain a good approximation to the optimal solution. In this respect, the convergence criterion checks if the last n_δ solutions to the approximate problems were all similar in the sense that the moving standard deviation of the last n_δ solutions must be small. This implies that over the last n_δ iterations of the algorithm, no new path has been added which drastically changed the solution. This does not guarantee that there does not exist such a path in the set of paths we have not sampled; however, the finer the grid, the less "weight" will be attributed to each single possible path, so that with reasonably fine grids there will not be big jumps in the convergence of the optimization procedure since single paths do not carry as much weight, especially if we sample enough. In proposition 8 below we state formally what we mean by "sampling enough paths".

Before doing this, two important remarks are warranted concerning the linear programs we solve in stage 2. First, note that in practice we introduce different errors into the linear programs. The main error is the "sampling error" introduced by estimating $\hat{F}_{Y,X|Z=z;h_n}$ in practice, which we do via the 'np'-package in R (Hayfield & Racine 2008) using a cross-validated bandwidth. In particular, the linear programs are likely to not admit a solution due to the equality constraint $\Theta\mu = \hat{F}_{Y,X|Z=z;h_n}$ in practice. This might even happen theoretically, but the practical approximation aggravates this issue. This is an important challenge we have to overcome in practice.

We deal with this issue by replacing the theoretical linear programs in stage 2 of the algorithm by their relaxed versions

$$\begin{array}{lll} \underset{\mu\geq 0,\, \vec{1}'\mu\leq 1}{\text{minimize}} & \left(\Xi_k^{min}\right)'\mu & \text{and} \\ \text{s.t.} \quad \|\Theta_k^{min}\mu - \hat{F}_{Y,X|Z=z;h_n}\|_2^2 \leq \varepsilon_{min} & & \end{array} \qquad \begin{array}{ll} \underset{\mu\geq 0,\, \vec{1}'\mu\leq 1}{\text{maximize}} & (\Xi_k^{max})'\mu \\ \text{s.t.} \quad \|\Theta_k^{max}\mu - \hat{F}_{Y,X|Z=z;h_n}\|_2^2 \leq \varepsilon_{max} & \end{array} \tag{13}$$

for some small $\varepsilon_{min}, \varepsilon_{max} > 0$. The above relaxed versions of the linear programs are equivalent to the penalized programs

$$\begin{array}{ll} \underset{\mu\geq 0, \vec{1}'\mu\leq 1}{\text{minimize}} & \left(\Xi_k^{min}\right)'\mu + \dfrac{\lambda_{min}}{2}\|\Theta_k^{min}\mu - \hat{F}_{Y,X|Z=z;h_n}\|_2^2 \qquad \text{and} \\ \underset{\mu\geq 0, \vec{1}'\mu\leq 1}{\text{maximize}} & (\Xi_k^{max})'\mu - \dfrac{\lambda_{max}}{2}\|\Theta_k^{max}\mu - \hat{F}_{Y,X|Z=z;h_n}\|_2^2 \end{array} \tag{14}$$

for some penalties λ_{min} and λ_{max}, which is a standard result from the theory of convex optimization.

Second, we need to show that the finite dimensional versions are consistent for our infinite

dimensional program as the dyadic approximation becomes finer. For this we write the programs (13) as

$$\begin{aligned} &\underset{\mu_j \in \mathcal{P}_j^*([0,1]^2)}{\text{minmize/maximize}} \int \Xi_j(y_j^*, x_{0,j}, z_j, u, v)\mu_j(dv, du) \\ &\frac{1}{m_j}\sum_{i=1}^{m_j}\left(\hat{F}_{Y,X|Z=z_i;h_n}(y_i, x_i) - \int \Theta_j(y_i, x_i, z_i, u, v)\mu_j(du, dv)\right)^2 < \frac{1}{m_j}\varepsilon' \coloneqq \varepsilon_j, \end{aligned} \tag{15}$$

were $\varepsilon' > 0$ is ε_{min} or ε_{max} depending on whether we minimize or maximize. Note that we weight the squared Euclidean norm by $\frac{1}{m_j}$, which does not change the optimization problem for fixed j.

Here, $\mathcal{P}_j^*([0,1]^2)$ is the space of probability measures satisfying the same assumptions as the measures in $\mathcal{P}^*([0,1]^2)$, but supported on the dyadic points of the corresponding dyadic approximation of order j. Ξ_j and Θ_j are the vectors as in (13) and the points y_j^* and $x_{0,j}$ are the dyadic points closest to but smaller than y^* and x_0; $\|\cdot\|$ is the Euclidean norm. The consistency of our finite dimensional programs is then captured in the following

Proposition 7. *Let Assumptions 1 – 4 hold. As the order j of the dyadic approximation increases to infinity the optimal solutions of the finite dimensional programs* (15)*, with $\hat{F}_{Y,X|Z;h_n}$ replaced by $F_{Y,X|Z}$, converge to optimal solutions of the relaxed infinite dimensional programs*

$$\begin{aligned} &\underset{\mu \in \mathcal{P}^*([0,1]^2)}{\textit{minimize/maximize}} \int \Xi(y^*, x_0, z, u, v)\mu(dv, du) \\ &\textit{s.t.} \quad \left\| F_{Y,X|Z=z}(y, x) - \int \Theta(y, x, z, u, v)\mu(du, dv)\right\|^2_{L^2([0,1]^3)} \leq \varepsilon \end{aligned} \tag{16}$$

for every $\varepsilon > 0$ and fixed $F_{Y,X|Z=z}$. $\|\cdot\|_{L^2([0,1]^3)}$ denotes the L^2-norm with respect to Lebesgue measure on $[0,1]^3$.

Proposition 7 shows that optimal solutions of the relaxed finite dimensional programs converge to optimal solutions of the infinite dimensional programs for continuous $F_{Y,X|Z}$, but it does not state that every optimal μ to the infinite dimensional programs (16) has a sequence of optimal measures that converge to them. This is not a problem in our setting, however, as we only care about the objective function of the optimization problems, for which we only need *one* optimal measure μ. We only prove the proposition for the relaxed programs as we use the relaxed finite dimensional programs in practice.

Having addressed the two comments, we can now formally state what we mean by "sampling enough" paths. First, note that we consider this question with respect to a given finite dyadic approximation, since in general we cannot hope to obtain a fine covering of an infinite dimensional set with finitely many paths. Second, the notion of "enough" should not be taken too literally. In particular, we cannot obtain results which show how far the optimal solution of the sampled problems is from the problems with all possible paths without making strong assumptions. What

we can do is to derive a probabilistic result which provides a lower bound on "enough" if we want to be reasonably confident that the optimization problem over the sampled finite dimensional linear program would also be optimal over the infinite dimensional program. For this, we introduce the dual problems of (14), which turn out to be second order polynomials in the dual variable y:

$$\min_{y\geq 0} \; -\frac{1}{2}y^2\vec{1}'\left(\lambda_{min}D_k^{min}\right)^{-1}\vec{1}+y\left[\vec{1}'\left(\lambda_{min}D_k^{min}\right)^{-1}\left[\lambda_{min}\left(\Theta_k^{min}\right)'\hat{F}_{Y,X|Z=z;h_n}-\Xi_k^{min}\right]-1\right]+R_{min}$$

$$\min_{y\geq 0} \; -\frac{1}{2}y^2\vec{1}'\left(\lambda_{max}D_k^{max}\right)^{-1}\vec{1}+y\left[\vec{1}'\left(\lambda_{max}D_k^{max}\right)^{-1}\left[\lambda_{max}\left(\Theta_k^{max}\right)'\hat{F}_{Y,X|Z=z;h_n}+\Xi_k^{max}\right]-1\right]+R_{max},$$

where $D_k^{min} \coloneqq \left(\Theta_k^{min}\right)'\Theta_k^{min}$ and $D_k^{max} \coloneqq \left(\Theta_k^{max}\right)'\Theta_k^{max}$, and where R_{min} and R_{max} are the constants of the second order polynomials. Since both problems are well-behaved, it follows that the optimal values of the primal problems coincide with the optimal values of the dual problems, so that we can use the latter to derive our result for the following sampling result, which is similar to the a sampling result in Pucci de Farias & Van Roy (2004).

Proposition 8 (Probabilistic finite sample near optimality)**.** *Fix a finite dyadic decomposition of the unit interval of order j, yielding m_j points. Sample, with some probability P_s, paths of the finite set $\mathcal{W}_j$ of all continuous paths supported on the m_j points. Then for every $\varepsilon,\delta\in(0,1)$ there exists a sample $\mathcal{W}(s(\delta,\varepsilon))$ drawn from $\mathcal{W}_j$ of at least size*

$$s(\delta,\varepsilon)\geq\frac{4}{\varepsilon}\left(3\ln\left(\frac{12}{\varepsilon}\right)+\ln\left(\frac{4}{\delta}\right)\right)$$

such that with confidence $1-\delta$ it holds that

$$\sup_{\{y:\, a_{w,max}y^2+yb_{w,max}+r_{w,max}\leq V_{w,max}\}} P_s\left(\left\{i: a_{i,max}y^2+yb_{i,max}+r_{i,max}>V_{w,max}\right\}\right)\leq\varepsilon \qquad \textit{and}$$

$$\sup_{\{y:\, a_{w,min}y^2+yb_{w,min}+r_{w,min}\geq V_{w,min}\}} P_s\left(\left\{i: a_{i,min}y^2+yb_{i,min}+r_{i,min}<V_{w,min}\right\}\right)\leq\varepsilon,$$

where

$$\begin{aligned}
a_{w,max} &\coloneqq \vec{1}'\left(\lambda_{max}\tilde{D}^{max}\right)^{-1}\vec{1},\\
b_{w,max} &\coloneqq \left[\vec{1}'\left(\lambda_{max}\tilde{D}^{max}\right)^{-1}\left[\lambda_{max}\left(\tilde{\Theta}^{max}\right)'\hat{F}_{Y,X|Z=z;h_n}-\tilde{\Xi}^{max}\right]-1\right],\\
r_{w,max} &\coloneqq \tilde{R}_{max},\\
a_{i,max} &\coloneqq \vec{1}'\left(\lambda_{max}\underline{D}^{max}\right)^{-1}\vec{1},\\
b_{i,max} &\coloneqq \left[\vec{1}'\left(\lambda_{max}\underline{D}^{max}\right)^{-1}\left[\lambda_{max}\left(\underline{\Theta}^{max}\right)'\hat{F}_{Y,X|Z=z;h_n}-\underline{\Xi}^{max}\right]-1\right],\\
r_{i,max} &\coloneqq \underline{R}_{max},
\end{aligned}$$

for $\tilde{D}^{max}$, $\tilde{\Theta}^{max}$, $\tilde{\Xi}^{max}$, *and* $\tilde{R}_{max}$ *which are made up of only paths in* $\mathcal{W}(s(\delta,\varepsilon))$ *and* $\underline{D}^{max}$, $\underline{\Theta}^{max}$, $\underline{\Xi}^{max}$, *and* $\underline{R}_{max}$ *which are made up of those paths who yield a greater value than the optimal value for the sampled linear program* $V_{w,max}$. *Analogously for the minimization.*

Three remarks concerning this result are in order. First, the sampling rate $O(\frac{4}{\varepsilon}\left(3\left(\frac{12}{\varepsilon}\right)+\left(\frac{4}{\delta}\right)\right))$ does *not* directly depend on m_j, a surprising result at first glance. This follows from the fact that the dual programs can be written as a univariate second order polynomial in y, and it is known that polynomials have finite VC-dimension (Vapnik & Chervonenkis 1971) which does not depend on the number of points m_j. Second, Proposition 8 gives a *theoretical* lower bound on how many paths we have to sample in order to be confident, with probability $1-\delta$, that the solution of the sampled linear program coincides with the optimal solution of the general linear program with probability $1-\varepsilon$, where the latter probability is measured with respect to the sampling measure P_s. Note that this result does not say anything about how far the optimal solution of the sampled linear program is from the general linear program. We only provide a probabilistic result which states that we can be confident that the optimal solution of our sampled linear program coincides with the general linear program.

Third, in practice one should sample as many paths as possible given the memory and time constraints and make sure that one samples all forms of admissible paths in order to span the space of all paths adequately. The reason is that even for small ε, the set

$$\left\{i : a_{i,max}y^2 + yb_{i,max} + r_{i,max} > V_{w,max}\right\}$$

will still consist of several million paths if our overall set of all paths for a given dyadic order j has a cardinality of several billion. Note that since the paths grow super-exponentially, this happens already for dyadic approximations of low order, so that even though the probability of sampling those paths with P_s is less than ε, those are still a very large number of possible paths. In practice one should always sample enough paths until one sees convergence of the optimal values. In our practical application in the next section we provide pictures which show what we mean by convergence of the optimal values.

To not make the memory requirements grow too fast, we add stage 3 in the algorithm, dropping paths which turn out to not have an influence on the optimal values. This approach of dropping paths which are not relevant in stage 3 is similar in spirit to the corresponding stage in the relaxed cutting plane approach introduced in Wu, Fang & Lin (2001) for solving the general capacity problem. Recall that μ is a finite sample approximation for a given dyadic approximation, so that zero-elements in the vector μ doe not satisfy the relaxed constraint, which is the main idea for discarding these paths.

In order to solve the programs (14) efficiently we apply the alternating direction method of multipliers (ADMM) (Boyd, Parikh, Chu, Peleato & Eckstein 2011 and Parikh & Boyd 2014) for quadratic programs. This algorithm is known to converge rather quickly to reasonable approximations of the optimum, which makes it a perfect tool for our purposes. For the algorithm we need

to specify two more parameters, the augmented Lagrangian parameter ρ and an over-relaxation parameter α, which control the convergence of the ADMM algorithm to the optimum. In practice, we found that an over-relaxation parameter of $\alpha = 1.7$ leads to quick convergence. For more information on the ADMM algorithm we refer to Boyd, Parikh, Chu, Peleato & Eckstein (2011) and Parikh & Boyd (2014).

6 Application: Estimating expenditures

We apply our algorithm to estimate bounds on causal effects of expenditures using the 1995/1996 UK family expenditure expenditure survey. Analogous to Blundell, Chen & Kristensen (2007) and Imbens & Newey (2009), the outcome of interest Y will be the share of expenditure on a commodity, in our case food or leisure, and X will be the log of total expenditure, scaled to lie in the unit interval. The instrument we use 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 and hence fit perfectly into our model. We use a subset of married and cohabiting couples where the head of the household is aged between 20 and 55, and couples with 3 or more children are excluded. We also exclude households where the head of the household is unemployed in order to have the instrument available for each observation. This leaves us with 1650 observations. The only structural assumption we make on the model is continuity, i.e. we assume that 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.

Figure 1 provides a typical example of the performance of our algorithm for a very coarse grid. It depicts the convergence of the upper and lower bound for the values $y^* = 0.75$ and $x_0 = 0.75$, where the outcome Y is the share of expenditure on leisure for different levels of the penalization parameters λ_{min} and λ_{max}, and where we have scaled X and Z to lie in the unit interval. Note that we use a very coarse dyadic approximation of order 2 of the unit interval for this performance analysis, giving us the 5 grid-points $(0, 0.25, 0.5, 0.75, 1)$. This allows us check if our algorithm actually converges when we are able to sample *all* possible paths with high probability—we sample 16 new paths each time, giving us $160,000$ paths overall, which are far more than the maximal number of continuous paths for this coarse dyadic approximation, so that with very high probability we managed to sample every possible path.

Note that the bounds are rather large, but not trivial.[22] The size of the bounds is to be expected due to the very coarse approximation; in actuality, we had expected completely trivial bounds due to the coarseness of the grid, so a non-trivial lower bound at this stage might suggest that the information-to-noise-ratio is rather high in the data to answer the respective question.

[22] They are non-trivial in all estimations we ran, not just this one.

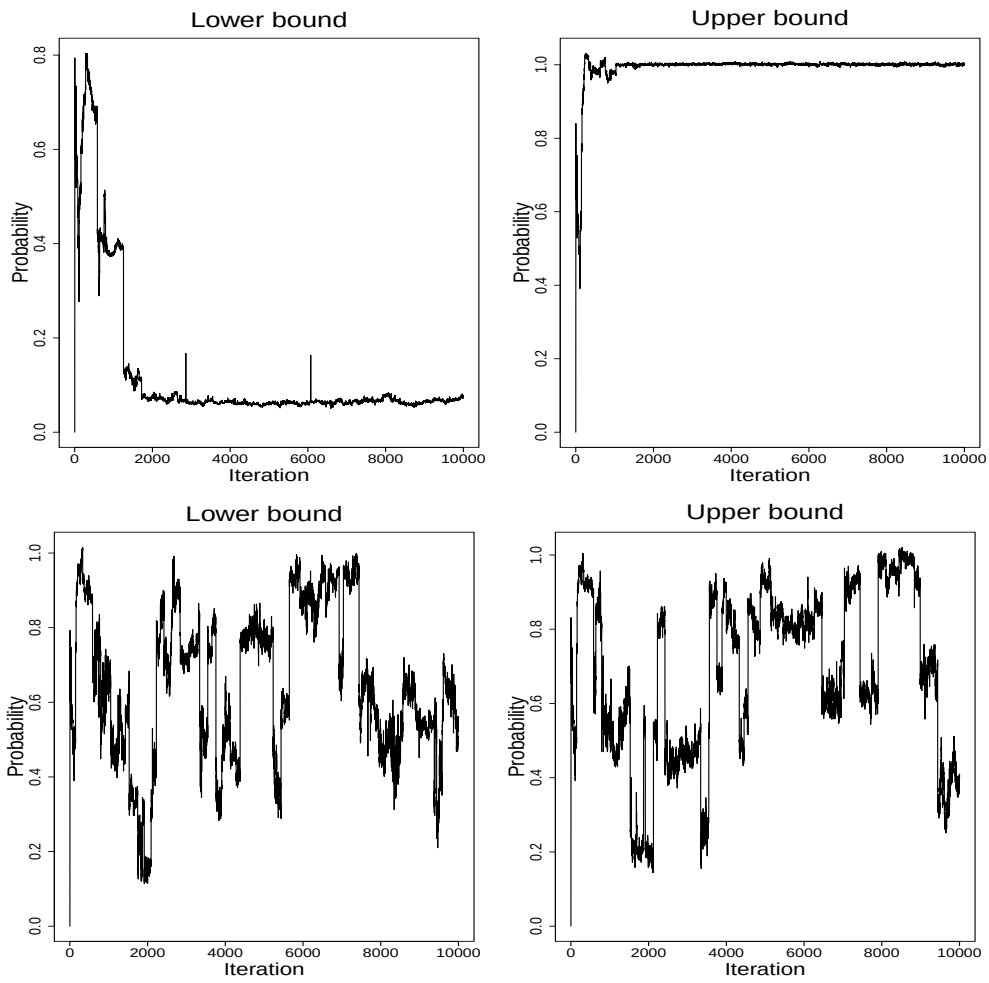

Figure 1: Convergence of upper and lower bounds on $F_{Y|X^*=0.75}(0.75)$ for $\lambda_{min} = \lambda_{max} = 100$ (top) and $\lambda_{min} = \lambda_{max} = 600$ (bottom) and a very coarse dyadic approximation of order 2; we sample 16 new paths at each iteration. For the lower value of $\lambda.$, we see a clear convergence of the algorithm, whereas the algorithm does not converge for the higher value of $\lambda.$.

Note the dependence on the penalty-terms λ_{min} and λ_{max}: for $\lambda_{min} = \lambda_{max} = 100$, we obtain convergence of the algorithm; the upper bound for $F_{Y|X^*=0.75}(0.75)$ is 1, and the lower bound is 0.065. As soon as we penalize the OLS-constraint too much, we do not obtain convergence of the algorithm. The solution paths of the programs depend on the penalization parameters λ_{min} and λ_{max}, which should be expected. In fact, if λ_{min} and λ_{max} are too small, we put almost no weight on the constraint. In contrast, if λ_{min} and λ_{max} are too big, we basically solve an OLS problem for the constraint and do not put any weight on the objective function, in which case the bounds will not converge in general, but only display erratic behavior, due to the sampling of the constraints approach. This again follows from the fact that the equality constraint does not admit a solution in general. We hence have a tradeoff for the penalty terms. These results only serve to check the convergence properties of our algorithm, but are too coarse to provide good

estimations of the bounds of the counterfactual probabilities.

The following results are for a more reasonable grid and constitute our main results for this application. Here, we use the dyadic approximation of order 4, yielding $m_4 = 17$ points on the unit interval. Figure 2 depicts the distributional causal effects for the lower quartile of the expenditure on leisure and food.

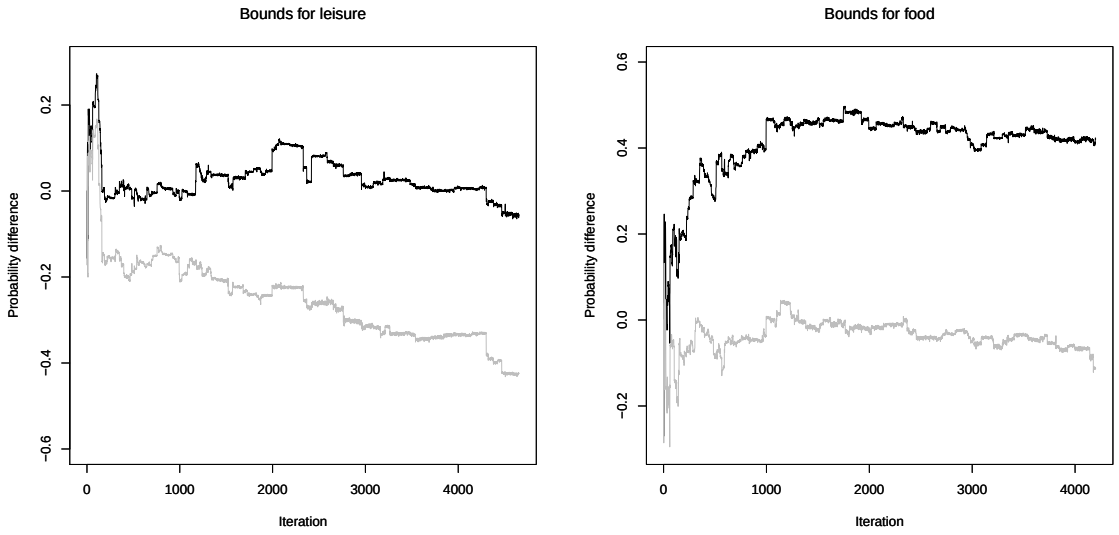

Figure 2: Convergence of upper (black) and lower (gray) bounds on $F_{Y|X^*=0.75}(0.25) - F_{Y|X^*=0.25}(0.25)$ for Y being the relative spending on leisure and food. The penalty terms are $\lambda_{min} = \lambda_{max} = 1$ and the dyadic approximation is of order 4. We sample 25 new paths at each iteration.

Here we had to lower the penalization parameters λ_{min} and λ_{max} to 1, because the grid is now much finer than before. The bounds are narrower than in the coarse approximation. Consider the left panel first. The upper bound is actually slightly negative for the last 300 iterations with -0.013 as an average while the lower bound is -0.39, providing very strong evidence that leisure is a luxury good. In fact, these bounds imply that families who do not spend a lot on leisure despite spending spend more overall (in the sense that they lie in the upper quartile of all families in overall spending) are very likely to spend even less on leisure, relatively, if they had a negative shock to overall spending. Put differently, families are much more likely to lie in the lower quartile for expenditure on leisure if they lie in the lower quartile in overall spending than families who lie in the upper quartile in overall spending, a strong indication that leisure is a luxury good. Note that the upper bound for leisure in Figure 2 does not have a trend and fluctuates between 0 and 0.05, which could be an indication that the theoretical upper bound is reached. In contrast, the lower bound for leisure has not leveled off, yet, implying that we would need to sample even more paths to get it to level off at some point. This implies that the actual lower bound is actually smaller than the one depicted here.

As for the right panel in Figure 2 the first thing to notice is that the bounds on the counterfactual probabilities are "shifted up" compared to the bounds on leisure. In fact, the lower bound is -0.025 while the upper bound is 0.4. This indicates that families who spend a lot in general and do not spend much on food compared to overall expenditure would spend much more on food relatively if they spent much less overall. 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 who lie in the upper quartile in overall spending. This is strong evidence for a necessity good.

Only considering the lower quartile for y^* for the difference $F_{Y|X^*=0.75}(y^*) - F_{Y|X^*=0.25}(y^*)$ does not provide enough evidence for our claim, however. Table 1 therefore provides the upper- and lower bounds for the expenditure on both leisure and food for different percentiles y^*.

	Leisure			Food		
y^*	# paths	Lower bound	Upper bound	# paths	Lower bound	Upper bound
0.05	108,225	-0.36	0.23	119,200	-0.088	0.27
0.15	121,250	-0.51	0.025	109,175	-0.18	0.39
0.25	116,375	-0.39	-0.031	105,000	-0.075	0.41
0.35	116,925	-0.16	0.043	108,175	-0.011	0.38
0.75	123,625	-0.051	0.021	117,125	-0.050	0.023

paths is the number of paths sampled, which is iterations $\times 25$.

Table 1: Upper- and lower bounds for $F_{Y|X^*=0.75}(y^*) - F_{Y|X^*=0.25}(y^*)$ for different percentiles y^*.

Table 1 provides "envelopes" for the counterfactual distributions and corroborates the findings from Figure 2 that food is a necessity while leisure is a luxury good by showing that the upper and lower bounds for food and leisure are rather similar for $y^* = \{0.15, 0.25, 0.35\}$. In fact, we see the clearest effects for lower probabilities. Figure 3 depicts the convergence of the distributional effects for $y^* = 0.35$. As y^* becomes bigger, the upper and lower bounds for both food and leisure become closer and centered around zero. This is expected as there are very few families who spend more than three quarters of their overall budget on food or leisure, so that we are not comparing many families for this quantile. In particular, for $y^* = 0.75$, there is basically no difference between families with a small budget and families with a large budget. Moreover, the convergence for higher values of y^*, in particular $y^* = 0.75$ is very smooth, another sign that there is not a lot of difference between families wit a large budget compared to families with a low budget in this echelon of spending on food and leisure. Figure 4 depicts this.

The results we have obtained are surprisingly clear. Recall that our model is completely general, so we did not make any assumptions during the estimation process. In particular, our model also incorporates measurement errors, which indicates that the ratio of information to noise in the data for answering these questions is rather high. Moreover, note that we did not restrict the heterogeneity in any way, despite the fact that U, V are univariate. In particular, the way our estimation problem is set up, the assumption of univariate U and V is without loss of generality as

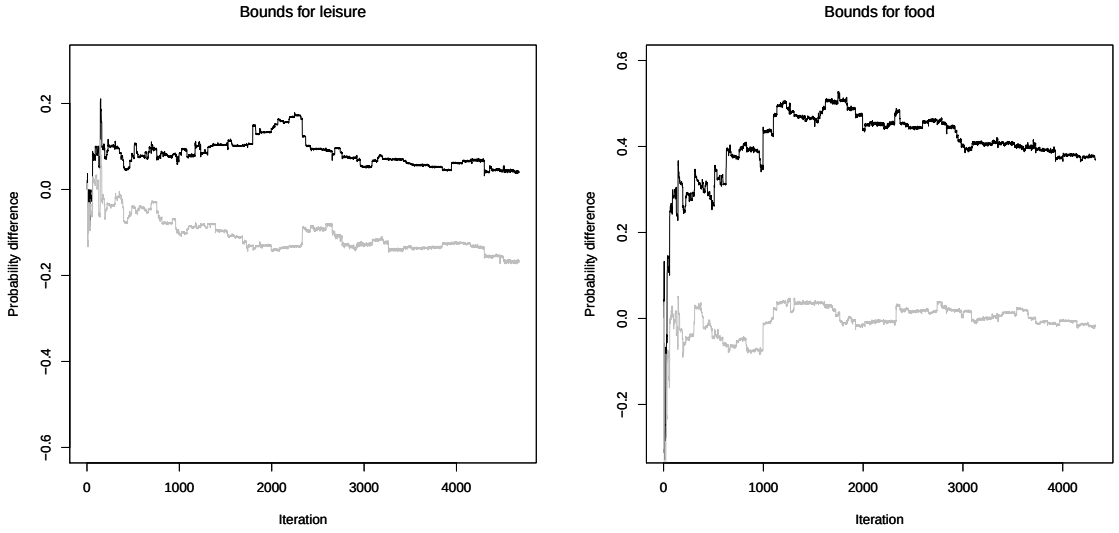

Figure 3: Convergence of upper (black) and lower (gray) bounds on $F_{Y|X^*=0.75}(0.35) - F_{Y|X^*=0.25}(0.35)$ for Y being the relative spending on leisure and food. The penalty terms are $\lambda_{min} = \lambda_{max} = 1$ and the dyadic approximation is of order 4. We sample 25 new paths at each iteration.

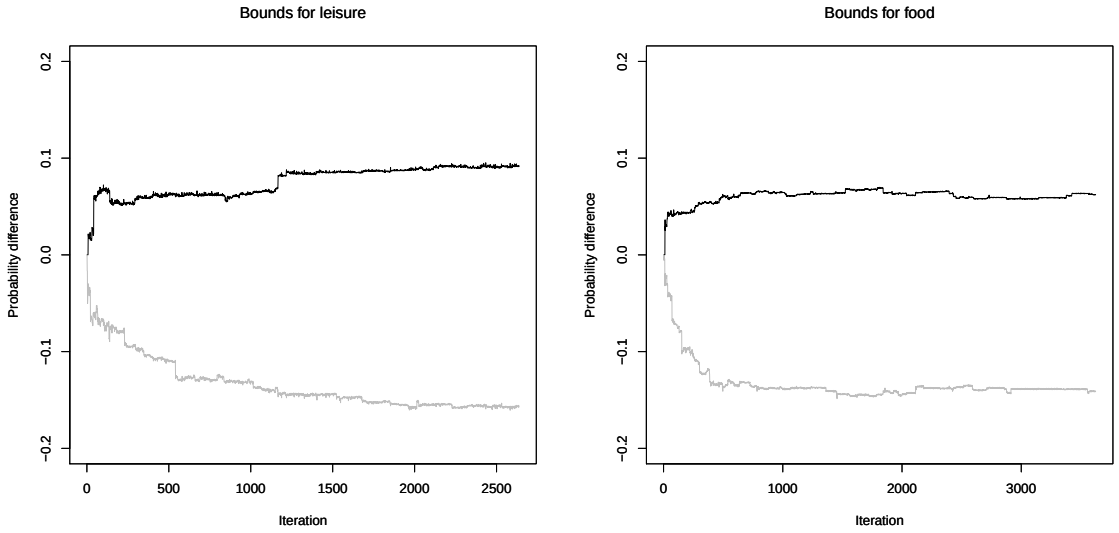

Figure 4: Convergence of upper (black) and lower (gray) bounds on $F_{Y|X^*=0.75}(0.75) - F_{Y|X^*=0.25}(0.75)$ for Y being the relative spending on leisure and food. The penalty terms are $\lambda_{min} = \lambda_{max} = 0.3$ and the dyadic approximation is of order 4. We sample 25 new paths at each iteration.

there are always (almost) invertible maps between the unit interval and any general (even infinite dimensional) spaces.

In fact, we only need some general measure space which can index all possible paths, and the

most convenient space for this is the unit interval. The difference to other identification results allowing for general unobserved heterogeneity is that we only work with the *cardinality* of the sets on which U and V are defined, because our identification problems are so general; all other results in the literature make additional topological or order-theoretic assumptions (invertibility of $g(Z,U)$ in U for instance), often restricting the dimension of the unobservables U and/or V. Simply using cardinality is the most general approach, which is why we can estimate models with the most general heterogeneity by restricting U and V to be the unit interval.

Overall, our results not only corroborate the theoretical predictions for expenditure, but also the previous results obtained in Blundell, Chen & Kristensen (2007), Imbens & Newey (2009), and Song (2018). Those articles estimate Engel curves for different commodities, in particular leisure and food; throughout they find decreasing Engel curves for food and increasing Engel curves for leisure, implying that leisure is a luxury while food is a necessity good. During their estimation process Imbens & Newey (2009) and Song (2018) assume a univariate and strictly monotonic production function $g(z,U)$ between X and U for all z and use a control variable approach to estimate the production function h; Blundell, Chen & Kristensen (2007) estimate Engel curves semi-nonparametrically, obtaining similar results. Our approach corroborates all of these results. In this sense our result is a "robustness check" for other non- or semiparametric approaches.

As a next step we want to check different identifying assumptions such as monotonicity directly in our general estimation procedure, gauging how strong the different identification assumptions are. Since those structural assumptions rule out many paths of stochastic processes, they should lead to tighter bounds all else equal. The size of these bounds then gives an indication for how strong those identifying assumptions are for the given problem at hand. We plan on pursuing this in a future article.

7 Conclusion

In this article we have analyzed two fundamental questions in instrumental variable models in the setting where the endogenous variable is continuous.

First, we have proved a slight generalization of the conjecture in Pearl (1995*b*), showing that the exclusion restriction of an instrument cannot be tested in general instrumental variable models with a continuous endogenous variable. The idea is to construct a general measure preserving isomorphism for the first stage, which is akin to the construction of an (almost everywhere) Condorcet cycle in uncountable state space. This result has several interesting implications for the general research on instrumental variable models. In particular, it implies that the continuous case is fundamentally different from the discrete case and that one should be cautious when arguing about testability of the continuous case by using discretization. Moreover, the construction of a Condorcet cycle implies that we need to allow for very general production functions in order to arrive at the impossibility theorem. This suggests that testability can be reestablished in this setting under some weak structural form assumptions.

Second, as our main result we have derived a new way to estimate sharp bounds on counterfactual probabilities, generalizing the approach in Balke & Pearl (1994) and Balke & Pearl (1997) to the continuous setting. The idea is to treat the two stages of the nonseparable triangular model as two general dependent stochastic processes, which enables us to derive the counterfactual probabilities by solving an infinite dimensional linear program over the paths. These programs allow for the complete nonparametric setting and in principle enable us to encompass any possible (nonparametric) functional form assumptions. In particular, this framework enables us to test the ability for identification of models of these functional form assumptions: stronger nonparametric functional form assumptions will lead to narrower bounds. We plan to explore this in a future paper. More fundamentally, it can serve as a new framework for general nonparametric identification in instrumental variable models.

One major challenge to overcome was the practical implementation of the infinite dimensional linear programs. We have done so by introducing a new "sampling of paths" approach, where we randomly draw paths of the respective stochastic processes and solve a finite dimensional linear program over these paths. We seem to be the first in the mathematical literature to introduce such an approach for solving general infinite dimensional linear programs over stochastic processes—a class of optimization problems which also include general stochastic optimal control problems. This approach works rather well and generates informative bounds even in very coarse approximations and no other assumptions than the continuity of the respective stochastic processes. We also show that optimal measures of the finite dimensional programs converge to optimal measures of the infinite dimensional programs if we let the dyadic approximation go to infinity, proving consistency of our finite dimensional approach.

The key in practice is to choose an appropriate penalization parameter for these problems which can be compared to other penalty parameters as in LASSO or ridge regression, only in a functional setting. Currently, we have no general rule for choosing this parameter, except for the rule of thumb that we should pick large penalty terms without losing convergence of the algorithm. We plan to explore statistical ways for choosing it in a future paper. Moreover, we want to optimize the sampling approach—in particular, sampling uniformly is almost surely not the optimal thing to do. We are currently testing other, more efficient sampling routines which cut down memory- and time requirements. We have written an R program which allows the user to apply our approach to any data-set as soon as all variables are normalized to the unit interval. We plan to extend this program into a full R package.

The bounds we obtain on the respective distributional causal effects on the relative expenditure on food and leisure corroborate the theoretical predictions from economic theory: we find that leisure is a luxury good whereas food is a necessity good, corroborating the predictions from economic theory. This is especially interesting since we did not make any functional form assumptions except continuity during the estimation process and the results we obtained are still clear. In a next step we plan on applying our approach to different and more refined areas of goods.

Our general framework has many other possible applications in a variety of fields and is the

natural generalization of the seminal complier-, defier-, always taker-, never taker distinction from Angrist, Imbens & Rubin (1996). In particular, it is well-suited to answer causal questions whenever the endogenous variable is continuous, which is the case in economics, finance, and medicine. The main advantage is that we directly solve the infinite dimensional programs in all generality. In general, researchers can also use our approach in an initial step in their research to gauge the "informational" content of the data or certain structural assumptions for the question at hand before taking a more structural or parametric estimation approach.

References

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

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.

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.

Anthony, M. & Biggs, N. (1997), *Computational learning theory*, Vol. 30, Cambridge University Press.

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

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

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

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

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

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

Bogachev, V. I. (2007), *Measure theory*, Vol. 2, 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.

Chang, J. T. & Pollard, D. (1997), 'Conditioning as disintegration', *Statistica Neerlandica* **51**(3), 287–317.

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.

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

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.

Fang, Z. & Santos, A. (2014), 'Inference on directionally differentiable functions', *arXiv: 1404.3763* .

Field, E., Pande, R., Papp, J. & Rigol, N. (2013), 'Does the classic microfinance model discourage entrepreneurship among the poor? Experimental evidence from India', *American Economic Review* **103**(6), 2196–2226.

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.

Gunsilius, F. (2018*a*), 'Point-identification in multivariate nonseparable triangular models', *arXiv: 1806.09680* .

Gunsilius, F. (2018*b*), 'Testability of the exclusion restriction in continuous instrumental variable models', *arXiv: 1806.09517* .

Halmos, P. R. (1956), *Lectures on ergodic theory*, Vol. 142, American Mathematical Society.

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

Heckman, J. J. (2001), 'Micro data, heterogeneity, and the evaluation of public policy: Nobel lecture', *Journal of Political Economy* **109**(4), 673–748.

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

Heckman, J. J. & Vytlacil, E. (2005), 'Structural equations, treatment effects, and econometric policy evaluation', *Econometrica* **73**(3), 669–738.

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

Hoderlein, S., Holzmann, H., Kasy, M. & Meister, A. (2017), 'Corrigendum: Instrumental variables with unrestricted heterogeneity and continuous treatment', *The Review of Economic Studies* **84**(2), 964–968.

Hoderlein, S., Holzmann, H. & Meister, A. (2017), 'The triangular model with random coefficients', *Journal of Econometrics* **201**(1), 144–169.

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

Imbens, G. W. (2007), 'Nonadditive models with endogenous regressors', *Econometric Society Monographs* **43**, 17.

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.

Imbens, G. W. & Wooldridge, J. M. (2009), 'Recent developments in the econometrics of program evaluation', *Journal of economic literature* **47**(1), 5–86.

Johnson, R. A. (1970), 'Atomic and nonatomic measures', *Proceedings of the American Mathematical Society* **25**(3), 650–655.

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.

Karlan, D. S. & Zinman, J. (2005), Elasticities of demand for consumer credit. Yale University working paper.

Karlan, D. & Zinman, J. (2011), 'Microcredit in theory and practice: Using randomized credit scoring for impact evaluation', *Science* **332**(6035), 1278–1284.

Kédagni, D. & Mourifié, I. (2015), Sharp instrumental inequalities: testing IV independence assumption. Pennsylvania State University and Toronto University working paper.

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

Kitagawa, T. (2010), Testing for instrument independence in the selection model. UCL 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* .

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

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

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

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

Manski, C. F. (2003), *Partial identification of probability distributions*, Springer Science & Business Media.

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

Matoušek, J. (2009), *Geometric discrepancy: An illustrated guide*, Vol. 18, Springer Science & Business Media.

Mendiondo, M. S. & Stockbridge, R. H. (1998), 'Approximation of infinite-dimensional linear programming problems which arise in stochastic control', *SIAM journal on control and optimization* **36**(4), 1448–1472.

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.

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

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

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

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

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

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

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

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

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.

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.

Vapnik, V. & Chervonenkis, A. Y. (1971), 'On the uniform convergence of relative frequencies of events to their probabilities', *Theory of Probability and its Applications* **16**(2), 264.

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 Notation

The paths of the stochastic processes we construct lie in different function spaces, depending on the assumptions we make. Our working assumption will require the paths to lie in the space of continuous functions on the unit interval, denoted by $C([0,1])$. We always equip $C([0,1])$ with the supremum norm $\|f\|_\infty \coloneqq \sup_{x\in[0,1]} |f(x)|$, which makes $C([0,1])$ a Banach space; in fact, under this norm $C([0,1])$ becomes a special Banach lattice, an AM space, under the corresponding partial order $f \leq g \in C([0,1])$ if and only if $f(x) \leq g(x)$ for all $x \in [0,1]$.

We say that some convex problem in general, and a linear program in particular, has *no duality gap* if solving the dual problem gives the same result as solving the original (primal) problem. For an overview of duality in infinite dimensional linear programs we refer to the monograph Anderson & Nash (1987) and for general convex programs to Rockafellar (1974). When working in general Banach spaces the standard duality bracket will be defined by $\langle f, \mu \rangle$.

The natural *dual space* of $C([0,1])$, $C^*([0,1])$, is the Banach lattice $\mathcal{M}([0,1])$ of all finite Borel measures on $[0,1]$, since $[0,1]$ is compact (Aliprantis & Border 2006, Corollary 14.15). The partial order $\leq$ on $\mathcal{M}([0,1])$ is defined by $\mu \leq \nu$ if and only if $\mu(A) \leq \nu(A)$ for all Borel sets $A \in [0,1]$. The subset $\mathcal{M}^+([0,1])$ of all positive Borel measures forms a convex cone in $\mathcal{M}([0,1])$. The

subset of all Borel probability measures on $[0,1]$, $\mathcal{P}([0,1])$, is the intersection of the unit sphere in $\mathcal{M}([0,1])$ with the positive convex cone $\mathcal{M}^+([0,1])$.

The space $\mathcal{M}([0,1])$ is *metrizable*. This is convenient as it enables standard arguments using sequences. We will also need different topologies on $\mathcal{M}([0,1])$. The first important topology is the weak topology induced by the notion of *weak convergence*. A sequence of measures $(P_n)_{n\in\mathbb{N}}$ converges to a measure P weakly on some support S, denoted by $P_n \Rightarrow P$, if and only if

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

for every bounded and continuous real function f on S (Billingsley 1999). The second, even weaker, topology we briefly mention is induced by the concept of *vague convergence* of measures. We say that the sequence $\{P_n\}$ converges vaguely to P on S if

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

for every continuous f which vanishes at infinity. The topology induced by vague convergence (the vague topology) is the weak* topology of $C([0,1])$. Since we work on the unit interval $[0,1]$ which is a compact space, vague and weak convergence coincide. The last topology we put on the space of Borel measures is induced by the *total variation norm* on $\mathcal{M}([0,1])$ defined by

$$\|P' - P\|_{TV} \coloneqq \sup_f \left(\int f(x)P'(dx) - \int f(x)P(dx) \right),$$

where f ranges over the set of all measurable functions from $[0,1] \to [-1,1]$. If P' and P are probability measures we can write the total variation distance as

$$\|P' - P\|_{TV} \coloneqq \sup_{A\in\mathscr{B}} |P'(A) - P(A)|,$$

where $\mathscr{B}$ is an appropriate Borel σ-algebra.

Under our working assumptions it will turn out that the paths of the stochastic processes $Y_x(v)$ and $X_z(u)$ are actually *Hölder continuous*, i.e. lie in some Hölder space. To introduce the notion of Hölder spaces, let $k \coloneqq (k_1, \ldots, k_d)$ be a multi-index of non-negative integers $k_1, \ldots, k_d$. Set $|k| = \sum_{i=1}^d k_i$ and for some function $f(x_1, \ldots, x_d)$ write

$$D^k f \coloneqq \frac{\partial^{|k|}}{(\partial x_1)^{k_1} \cdots (\partial x_d)^{k_d}} f(x_1, \ldots, x_d),$$

which is a general form of writing the partial-derivatives-tensors of a multivariate function. Based

on this we define the *Hölder norm* of f by

$$\|f\|_{C^{k,\lambda}} \coloneqq \sum_{0\leq|a|\leq k} \|D^a f\|_\infty + \sum_{|a|=k} \sup_{x_1,x_2\in[0,1]:x_1\neq x_2} \frac{\|D^a f(x_1) - D^a f(x_2)\|}{|x_1 - x_2|^{(\lambda-\lfloor\lambda\rfloor)}}.$$

$\|\cdot\|$ stands for some norm in $\mathbb{R}^d$. The Hölder space $C^{k,\lambda}$ then consists of all functions with finite Hölder norm, i.e. $f \in C^{k,\lambda}([0,1])$ if and only if $\|f\|_{C^{k,\lambda}} < \infty$. For the construction of the respective stochastic processes, we also use the closed subsets $C^{0,\lambda}_K([0,1]) \coloneqq \{f \in C^{0,\lambda}([0,1]) : \|f\|_{C^{0,\lambda}} \leq K\}$ of $C^{0,\lambda}([0,1])$. We call a function *smooth* if it lies in all Hölder spaces, i.e. if $C^\infty([0,1])$.

In addition to standard functions, we also briefly work with the Dirac-delta distribution $\delta(x)$, which is the point-evaluation functional on the space of smooth functions with compact supports. The Dirac-delta distribution cannot be represented as a function as it satisfies $\int \delta(x-x_0)f(x)dx = f(x_0)$ as well as $\int \delta(x)dx = 1$. We will call $\delta(x)$ a distribution throughout, since we think it will not be confused with the concept of a probability distribution function.

Let us now turn to more measure theoretic concepts. As in the previous section we define a measurable space on some space Ω by $(\Omega, \mathscr{S})$, where $\mathscr{S}$ is the corresponding σ-algebra. We denote Borel σ-algebras by $\mathscr{B}$. Once we put a measure P on the measurable space, it becomes a measure space $(\Omega, \mathscr{S}, P)$. A measure space is *complete* if every subset of a measure zero set is also of measure zero. Note that the Borel σ-algebra is not complete. Its completion is the Lebesgue σ-algebra, which we will denote by $\mathscr{A}$. The cardinality of $\mathscr{A}$ is that of the power set of the continuum, whereas the cardinality of the Borel σ-algebra is that of the continuum, i.e. is smaller.

A random variable X is a measurable function $X : \Omega \to \mathbb{R}^d$, $d \geq 1$, which *pushes forward* the measure P defined on Ω to a measure P_X on $\mathbb{R}^d$ by $P_X(E_x) = P(X^{-1}(E_x))$, where $E_y \in \mathscr{B}_{[0,1]^d}$ and X^{-1} denotes the preimage of X as defined in section 3. The measure P_X is called the pushforward measure of P through X. By π_i we define the generic projections onto the i-th coordinate of some vector space. For example, for a vector $x \in \mathbb{R}^d$ $\pi_i x = x_i$. For a measure μ supported on $\mathbb{R}^d$ we mean the marginal distribution on the i-th coordinate after integrating out all other coordinates when writing $\pi_i\mu$. To avoid confusion with the smoothness parameter λ defined above, we denote Lebesgue measure on $\mathbb{R}^d$ simply by dx. Moreover, following standard mathematical conventions, we denote the linear integral operator

$$Tf(x) \coloneqq \int T(x,y)f(y)dy$$

and its kernel $T(x,y)$ by the same letter.

We also need to introduce ways to construct stochastic processes. For this we work with *dyadic intervals* in $[0,1]$, i.e. intervals with endpoints of the form $\frac{j}{2^m}$, where $j, m \in \mathbb{N}$. Note that the dyadic numbers form a dense subset of the real number line which is important to construct continuous stochastic processes. A dyadic interval Q in $\mathbb{R}^k$ is of the form $Q \coloneqq [a_1, b_i] \times [a_2, b_2] \times \cdots \times [a_k, b_k]$

and we define its length by $l(Q) = \max_{i\in\{1,\dots,k\}} |b_i - a_i|$. $\mathbb{I}$ denotes the set of irrational numbers in $[0,1]$ and will be identified with the *Baire space* $\mathbb{N}^{\mathbb{N}}$ (Aliprantis & Border 2006, 3.14). The Baire space is the space of all sequences of natural numbers and is helpful to assign unique "postal codes" to the irrational and hence real numbers.

To do this, we define for some finite vector of natural numbers $(n_1,\dots,n_k) \in \mathbb{N}^k$,

$$\langle n_1,\dots,n_k\rangle \coloneqq \{(m_i)_{i\in\mathbb{N}} \in \mathbb{N}^{\mathbb{N}} : m_1 = n_1,\dots,m_k = n_k\}$$

the identification of this vector with a set in Baire space. In particular, this identification is the set of all infinite postal codes in Baire space which start with the numbers $n_1,\dots,n_k$. By $\mathbb{N}^*$ we denote the set of all n-tuples of natural numbers, i.e. $\mathbb{N}^* \coloneqq \bigcup_{k=1}^{\infty} \mathbb{N}^k$. For more information on the Baire space we refer to Aliprantis & Border (2006, Chapter 3.14).

Based on the notion of an interval, we define a *quasi-interval* $I(n;t_1,\dots,t_n;a_1,b_1;\dots;a_n,b_n)$ as a subset of $\mathbb{R}^{[0,1]}$ of the form

$$I(n;t_1,\dots,t_n;a_1,b_1;\dots;a_n,b_n) \coloneqq \{f \in \mathbb{R}^{[0,1]} : a_i \le f(t_i) \le b_i, i = 1\dots,n\}.$$

This is a set of some intervals in $\mathbb{R}$ defined at respective points $t_1,\dots,t_n \in [0,1]$ and is needed for constructing stochastic processes. Intuitively, quasi-intervals are needed to define all real-valued stochastic processes in $\mathbb{R}$ which pass through the interval $[a_i,b_i]$ at time t_i. Quasi-intervals are a special type of *cylinder sets*

$$I_{t_1,\dots,t_k}(B) \coloneqq \{f \in \mathbb{R}^{[0,1]} : (f(t_1),\dots,f(t_k)) \in B, B \in \mathscr{B}_{\mathbb{R}^k}\}.$$

$I_{t_1,\dots,t_{2^m}}(B)$ is a *dyadic quasi-interval* of order m provided that $t_j = \frac{j}{2^m}$, $j = 0,1,\dots,2^m$ and B is a dyadic interval in $\mathbb{R}^{2^m}$ of length $l(B) \le \frac{1}{2^m}$. On each quasi-interval $I_{t_1,\dots,t_k}(B)$ one can define a "size" by $P(I_{t_1,\dots,t_k}(B))$ through some finite measure P.

For the construction of the respective stochastic processes we also use F_σ sets, which are uncountable unions of closed sets. In this respect we denote the interior of a set A by A°, its closure by $\bar{A}$, and its complement by A^c. $\bigcup\!\!\!\cdot$ denotes disjoint union of sets. A map $\phi : X \to Y$ between measure spaces equipped with the Borel σ-algebra is a *Borel*-isomorphism if ϕ and ϕ^{-1} are Borel measurable. Following the definitions of Kuratowski (1934) and Hess (1982), ϕ is a $(0,1)$-homeomorphism if $\phi^{-1}(B)$ is open and $\phi(A)$ is an F_σ set for open $B \subset Y$ and $A \subset X$; it is called a $(1,1)$-homeomorphism if $\phi^{-1}(B)$ and $\phi(A)$ are both F_σ sets. Lastly, we call a set pre-compact if its closure is compact.

B Proofs omitted from the main text

B.1 Proof of Lemma 1

Proof. Let $g(Z,U)$ be a one-to-one generator of the measure $P_{X|Z}$ and factor $P_{Y,X|Z} = P_{Y|X,Z}P_{X|Z}$. Use $x = g(z,u)$ to generate $P_{X|Z=z}$ via P_U and some other function $y = h'(x,z,v')$ to generate $P_{Y|X=x,Z=z}$ via $P_{V'}$, where u and v' are independent. Note that h' exists by Theorem 9.2.2 in Bogachev (2007). Since g is one-to-one in z, one can invert it to obtain $z = g^{-1}(x,u)$ and substitute this into h' to get $y = h'(x, g^{-1}(x,u), v') = h(x,u,v')$, which conforms to Model (1) if we consider (u,v') as v. Note that since U and V' are independent and both can be made independent of Z by changing the unobservable production functions h' and g, it holds that $Z \perp\!\!\!\perp (V,U)$. This construction can always be achieved, even if we require U and V to be univariate, as every Polish space equipped with a Borel probability measure is isomorphic to the unit interval with some probability measure (Bogachev 2007, Theorem 9.2.2), so that there always exists a probability measure for V which is the pushforward of a probability measure for (U,V'). □

To show in a simple example that restrictions on the dimension of U and V cannot help, assume that all random variables take values in the unit interval $[0,1]$. It then holds that $(u,v') \in [0,1]^2$ while $v \in [0,1]$. In this case one can construct a standard Hilbert- or Peano curve $\mathscr{H} : [0,1] \to [0,1]^2$ which is a measure preserving isomorphism from the unit interval to the unit square. Theorem 9.2.2 in Bogachev (2007) shows that this is a special case of a more general property, which is the property we need for the impossibility result.

B.2 Proof of Pearl's conjecture

For the proof of the conjecture we need to introduce two additional mathematical concepts besides *generators*: *measure preserving isomorphisms*, and *disintegrations*. The concept of disintegrations gives meaning to the restriction of a joint probability measure $P_{Y,X}$ to a subset of Lebesgue measure zero, for instance the conditional measure $P_{Y|X=x}$ when X is a continuous random variable inducing a nonatomic probability measure P_X. A disintegration $P_{Y|X=x}(A)$ for some Borel set A is a version of the standard conditional expectation $E(\mathbb{1}\{Y \in A\}|\mathscr{F}_X)$ for some filtration $\mathscr{F}_X \subset \mathscr{B}_X$ when it exists, where $\mathbb{1}\{E\}$ denotes the standard indicator function which is 1 if the event E happens and 0 otherwise. The existence of a disintegration can be shown under very general circumstances and is guaranteed in our setting (see Theorem 1 in Chang & Pollard 1997). Working with disintegrations instead of general conditional expectations makes the proof less burdensome in terms of notation.

The second formal concept we require for the proof of the conjecture is the notion of measure preserving isomorphisms. A map $T : \mathcal{X} \to \mathcal{Y}$ transporting a probability measure P_X onto another

probability measure P_Y is *measure preserving* if it is measurable[23] and

$$P_Y(E) = P_X(T^{-1}E) \tag{17}$$

for every set E in the Borel σ-algebra $\mathscr{B}_Y$ corresponding to Y.[24] If T is invertible and its inverse is also measure preserving, it is called a measure-preserving isomorphism. For all our work, we only need measure preserving isomorphisms up to sets of measure zero, as measure preserving isomorphisms can only be identified up to sets of measure zero anyways. Therefore, from now on we mean "measure preserving isomorphism modulus sets of measure zero" when we write "measure preserving isomorphism".

We can now turn towards the proof of the conjecture. As mentioned, the key is Lemma 1, which reduces the problem of proving the conjecture to simply proving that there exists a one-to-one generator for each possible $P_{X|Z}$. Also note that by using Lemma 1 we do not need to make any assumptions on the distribution of Y, so that we can allow for general distributions here too, which does not change the result. In fact, the whole proof works with the properties of $P_{X|Z=z}$, which we assume to be nonatomic for almost all $z \in \mathcal{Z}$.

For the construction of the Condorcet cycle in our proof, we need the following technical lemma about measure preserving isomorphisms, which is proved in Halmos (1956, p. 74).

Lemma 5. *Fix some probability measure m on a measurable space $(\mathcal{X}, \mathscr{B}_\mathcal{X})$, where $\mathcal{X} \subset \mathbb{R}$ is an interval. If E and F are Borel sets of the same measure in the interval $\mathcal{X}$, i.e. $m(F) = m(E)$, then there exists a measure preserving isomorphism $T : \mathcal{X} \to \mathcal{X}$ such that $m(TE + F) = 0$ on $\mathcal{X}$.*[25]

We are now ready to prove the conjecture. Note that we prove a slightly stronger statement than Conjecture 3, as we can construct some g which is a measure preserving isomorphism in both U and Z.

Proof of Pearl's conjecture. We will assume that all random variables take values in the unit interval, which is without loss of generality as for every probability measure μ on a Polish space there exists a measure preserving isomorphism onto the unit interval equipped with some probability measure ν; in case μ is nonatomic one can pick ν to be Lebesgue measure (Bogachev 2007, Theorem 9.2.2).

To begin notice that $g(z, \cdot)$ in model (1) is by definition a measure-preserving map $g(z, \cdot) : [0, 1] \to [0, 1]$ for almost every z. In fact, since $P_{X|Z=z}$ for almost all z and P_U are probability measures, we can require g to be a measure preserving isomorphism, i.e. to be invertible with measure preserving inverse by Theorem 9.2.2 in Bogachev (2007). Note also that g is only specified

[23] Measurability of T means that $\mathscr{B}_X = T^{-1}\mathscr{B}_Y$, where $\mathscr{B}_Y$ and $\mathscr{B}_X$ are the Borel σ-algebras corresponding to Y and X, respectively.

[24] $T^{-1}E$ denotes the set of points $x \in \mathcal{X}$ such that $Tx \in E$.

[25] For two sets A and B, $A + B$ denotes their disjoint sum $(A \setminus B) \cup (B \setminus A)$.

up to sets of measure zero, so that in our proof we only have to specify it modulus sets of measure zero, too. Of course, if we can prove the conjecture when choosing g to be a measure preserving isomorphism, we have also proved it for general measure preserving maps.

We proceed by first proving the conjecture for nonatomic P_Z with support[26] $\mathcal{Z} \subset [0,1]$. After that we extend the result to allow for P_Z with atoms.[27] In light of Lemma 1, we only have to show that $P_{X|Z=z}$ admits a one-to-one generator modulus sets of measure zero. We hence need to show that for each $z \in [0,1]$ there exists a measure preserving isomorphism $g(z,u)$ between P_U and $P_{X|Z=z}$ such that $g(z_i,u) \neq g(z_j,u)$ for almost all $z_i, z_j \in [0,1]$, $z_i \neq z_j$ and $u \in [0,1]$.

We show that such g exists by factoring it as $g(z,u) = T_z f_z(u)$. Here, $f_z : [0,1] \to [0,1]$ is some measure preserving isomorphism between P_U and the respective $P_{X|Z=z}$. We require T_z to be a measure preserving isomorphism on the measure space $([0,1], \mathscr{B}_{[0,1]}, P_{X|Z=z})$ garbling the map f_z.[28] To be more precise, since f_z is allowed to be *any* measure preserving isomorphism, it may well happen that $f_{z_i}(u) = f_{z_j}(u)$ for some $z_i, z_j, u \in [0,1]$. We therefore need to show that there always exists a collection T_z of measure preserving isomorphisms such that $T_{z_i} f_{z_i}(u) \neq T_{z_j} f_{z_j}(u)$ for almost all z_i, z_j, and $u \in [0,1]$. The idea to achieve this is to prove a simple generalization of the Condorcet Paradox to uncountable state space as outlined in the main text. The proof is therefore complete if we can show that a Condorcet cycle can always be constructed in the case where $\mathcal{X} = [0,1] = \mathcal{Z} = \mathcal{U}$ when T_z is a measure preserving isomorphism.

To show this, let P_U be Lebesgue measure on $[0,1]$ and let $f_z(u)$ be any measure preserving isomorphism from P_U to $P_{X|Z=z}$ for all z.[29] If $f_z(u)$ is a one-to-one generator, we just let T_z be the identity for all z and the conclusion follows. So assume that there are Borel sets $E_u \subset [0,1]$ and E_Z of measure $P_U(E_u) = \varepsilon_u$ and $P_Z(E_Z) = \varepsilon_Z$ for some $\varepsilon_u, \varepsilon_Z > 0$ such that $f_{z_i}(u) = f_{z_j}(u)$ for almost all $u \in E_u$ and $z_i, z_j \in E_Z$. Then we can define a measure preserving permutation $T_z : [0,1] \to [0,1]$ which is garbling of f_z in the sense that $T_{z_i} f_{z_i}(u) \neq T_{z_i} f_{z_j}(u)$ almost all $u \in E_u$ and $z_i, z_j \in E_Z$.

To do this, partition $E_Z = E_Z^1 \cup E_Z^2$ into two disjoint parts E_Z^1 and E_Z^2 of equal measure

$$P_Z(E_Z^1) = P_Z(E_Z^2) = \frac{1}{2}\varepsilon_Z$$

[26]The support $\mathcal{Z}$ of a measure P_Z is defined by two criteria. First, it is the closed set $\mathcal{Z}$ on which the measure P_Z concentrates, i.e. $P_Z([0,1] \setminus \mathcal{Z}) = 0$. Second, it is such that for every open set G such that $G \cap \mathcal{Z} \neq \emptyset$, it holds that $P(G \cap \mathcal{Z}) > 0$, i.e. every open set intersecting the support has positive measure. This is a straightforward generalization of the support of a density function and the reader can always think about the latter. It is also important to note that every probability measure on $\mathbb{R}^d$ has a support.

[27]Note that this covers all types of probability measures with finitely many atoms, even continuous singular ones on $\mathbb{R}$ as a general finite measure can be uniquely decomposed into a nonatomic measure and a purely atomic measure under the assumption that they are singular with respect to one another, see Theorem 2.1 in Johnson (1970). In fact, continuous singular measures like Cantor measures are nonatomic measures with a set of Hausdorff dimension smaller than one as support.

[28]Note that T_z is a measure preserving isomorphism which maps *to the same* measure space, whereas f_z maps between different measure spaces. We need this set-up since we want to use Lemma 5, which works for isomorphisms acting on the same measure space.

[29]We assume U to follow the uniform distribution on $[0,1]$ for convenience, we could specify any other distribution; also note that this is without any loss of generality, as we are free to choose the distribution of U for this proof.

and do the same with E_u, i.e. $E_u = E_u^1 \cup E_u^2$ with

$$P_U(E_u^1) = P_U(E_u^2) = \frac{1}{2}\varepsilon_u.$$

Since f_z is a measure preserving isomorphism for every $z \in [0,1]$, it must be the case that $f_z(E_u^1)$ and $f_z(E_u^2)$ are disjoint modulus sets of measure zero and that

$$P_{X|Z=z}(f_z(E_u^1)) = P_{X|Z=z}(f_z(E_u^2)) = \frac{1}{2}\varepsilon_u \quad \text{for all } z \in E_Z.$$

Define T_z to be the identity for $z \in E_Z^1$; for $z \in E_Z^2$ let it be such that $T_z f_z(E_u^1) = f_z(E_u^2)$ and $T_z f_z(E_u^2) = f_z(E_u^1)$, i.e. switching $f_z(E_u^1)$ and $f_z(E_u^2)$. Lemma 5 guarantees that this is always possible. Now partition E_Z^2 into two parts of equal measure E_Z^{21} and E_Z^{22}, so that

$$P_Z(E_Z^{21}) = P_Z(E_Z^{22}) = \frac{1}{4}\varepsilon_Z$$

and do the same with E_u^2. Then on E_Z^{21}, let T_z be the same as on E_Z^2 and on E_Z^{22} let it be such that $T_z f_z(E_u^{21}) = f_z(E_u^{22})$ and $T_z f_z(E_u^{22}) = f_z(E_u^{21})$.

At stage $n \in \mathbb{N}$ with sequences E_u^i and E_Z^i for $i \in \{1,2\}^n$, the inductive step is to split E_u^i and E_Z^i into two disjoint Borel subsets of equal measure $E_u^{i\wedge 1}$ and $E_u^{i\wedge 2}$ as well as $E_Z^{i\wedge 1}$ and $E_Z^{i\wedge 2}$.[30] Then on $E_Z^{i\wedge 1}$ let T_z be identical to T_z on E_Z^i and on $E_Z^{i\wedge 2}$ let it be such that

$$T_z f_z(E_u^{i\wedge 1}) = T_z f_z(E_u^{i\wedge 2}) \quad \text{and} \quad T_z f_z(E_u^{i\wedge 2}) = T_z f_z(E_u^{i\wedge 1}),$$

which is possible by Lemma 5. Now let us order the set $\{1,2\}^{\mathbb{N}}$ as follows: Start with the sequence of all ones, which is the minimal element. Then change the first digit from a 1 to a 2, keeping all other digits. Then change the first digit back to 1 and the second to a 2, keeping all others as 1. Let the digit 2 "run through all positions" up to infinity. After this change the first two digits to a 2, keeping all other digits at 1, and let the second 2 run through all positions, keeping the first position at 2. Change the first position back to 1 and keep the second position a 2 while running the second 2 through all positions. Do the same with three 2's, four 2's and so on. This is a well-ordering since every subset of $\{1,2\}^{\mathbb{N}}$ has a smallest element (Aliprantis & Border 2006, p. 18). Therefore, we can proceed by transfinite induction for all $i \in \{1,2\}^{\mathbb{N}}$ over this well-ordered set, which yields an uncountable Condorcet cycle for $T_z f_z(u)$ modulus sets of measure zero in the sense that $T_{z_i} f_{z_i}(u) \neq T_{z_i} f_{z_j}(u)$ almost all $u \in E_u$ and $z_i, z_j \in E_Z$.

This construction only uses values within E_Z and E_u, respectively, and can hence be applied separately to every combination of Borel sets E_Z and E_u for which $f_{z_i}(u) = f_{z_j}(u)$ for $u \in E_u$, yielding a Condorcet cycle up to sets of measure zero for all of $[0,1]$ if we let T_z be the identity for all other Borel sets. This proves the conjecture in the case where E_Z is uncountable since g

[30] The notation $i \wedge 1$ means appending the number 1 to the sequence i.

can only be specified modulus sets of measure zero.

For the case where Z has finitely many atoms, the above construction can be adapted as follows. If there is no atomic $z \in [0,1]$ such that $f_z(u) = f_{z_i}(u)$ for some other $z_i \in [0,1]$ and some $u \in [0,1]$, then the same construction as above works. If there is some number of atomic z_j, $j = \{1, \ldots, k\}$, for which there is a Borel set E_u and some Borel set E_Z such that $f_{z_j}(u) = f_{z_i}(u)$ for all $z_i \in E_Z$ and all $u \in E_u$, then in the construction above can be adjusted as follows.

Consider the Borel set $F \coloneqq \bigcup_{j=1}^{k} z_j \cup E_Z$ and partition E_Z into two disjoint Borel subsets E_Z^1 and E_Z^2 of equal measure

$$P_Z(E_Z^1) = P_Z(E_Z^2) = \frac{1}{2}\varepsilon_Z$$

and the corresponding E_u into $k+2$ disjoint Borel sets $E_u^1, \ldots, E_u^{k+2}$ of equal measure, i.e.

$$P_U(E_u^1) = \ldots = P_U(E_u^{k+2}) = \frac{1}{k+2}\varepsilon_u.$$

Then for z_1 let T_{z_1} be the identity. For the other values $z_2, \ldots, z_k$ as well as any $z \in E_Z^1$ and $z' \in E_Z^2$ let T_{z_j} be a cyclic map (which can be done by Lemma 5), i.e. for z_2

$$T_{z_2} f_{z_2}(E_u^{k+2}) = f_{z_2}(E_u^1), \quad T_{z_2} f_{z_2}(E_u^1) = f_{z_2}(E_u^2), \quad \ldots, \quad T_{z_2} f_{z_2}(E_u^{k+1}) = f_{z_2}(E_u^{k+2}),$$

for z_3

$$T_{z_3} f_{z_3}(E_u^{k+1}) = f_{z_2}(E_u^1), \quad T_{z_3} f_{z_3}(E_u^{k+2}) = f_{z_3}(E_u^2), \quad \ldots, \quad T_{z_2} f_{z_2}(E_u^{k}) = f_{z_2}(E_u^{k+2}),$$

for z_k

$$T_{z_k} f_{z_k}(E_u^1) = f_{z_k}(E_u^k), \quad T_{z_k} f_{z_k}(E_u^2) = f_{z_k}(E_u^{k+1}), \ldots,$$

for $z \in E_Z^1$

$$T_z f_z(E_u^1) = f_z(E_u^{k+1}), \quad T_z f_z(E_u^2) = f_z(E_u^{k+2}), \ldots,$$

and for $z' \in E_Z^2$

$$T_{z'} f_{z'}(E_u^1) = f_{z'}(E_u^{k+2}), \quad T_{z'} f_{z'}(E_u^2) = f_{z'}(E_u^1), \ldots$$

Then at each iteration n of the construction split $E_Z^{i_z}$, $i_z \in \{1,2\}^n$, into two disjoint Borel subsets $E_Z^{i_z \wedge 1}$ and $E_Z^{i_z \wedge 2}$ of equal measure

$$P_Z(E_Z^{i_z \wedge 1}) = P_Z(E_Z^{i_z \wedge 2}) = \frac{1}{2^n}\varepsilon_Z$$

and $E_u^{i_u}$, $i_u \in \{1, \ldots, k+2\}^n$ into $k+2$ disjoint Borel subsets $E_u^{i_u \wedge 1}, \ldots, E_u^{i_u \wedge k+2}$ of equal measure

$$P_U(E_u^{i_u \wedge 1}) = \ldots = P_U(E_u^{i_u \wedge k+2}) = \frac{1}{(k+2)^n}\varepsilon_u,$$

leave T_{z_1} unchanged from period $n-1$ on $E^{i_z \wedge 1}$, and construct it to be cyclic (which again can be done by Lemma 5) for the other z, i.e. for z_2

$$T_{z_2} f_{z_2}(E_u^{i_u \wedge k+2}) = f_{z_2}(E_u^{i_u \wedge 1}), \quad \ldots, \quad T_{z_2} f_{z_2}(E_u^{i_u \wedge k+1}) = f_{z_2}(E_u^{i_u \wedge k+2}),$$

for z_3

$$T_{z_3} f_{z_3}(E_u^{i_u \wedge k+1}) = f_{z_2}(E_u^{i_u \wedge 1}), \quad \ldots, \quad T_{z_2} f_{z_2}(E_u^{i_u \wedge k}) = f_{z_2}(E_u^{i_u \wedge k+2}),$$

for z_k

$$T_{z_k} f_{z_k}(E_u^{i_u \wedge 1}) = f_{z_k}(E_u^{i_u \wedge k}), \quad T_{z_k} f_{z_k}(E_u^{i_u \wedge 2}) = f_{z_k}(E_u^{i_u \wedge k+1}), \ldots,$$

for $z \in E_Z^{i_z \wedge 1}$

$$T_z f_z(E_u^{i_u \wedge 1}) = f_z(E_u^{i_u \wedge k+1}), \quad T_z f_z(E_u^{i_u \wedge 2}) = f_z(E_u^{i_u \wedge k+2}), \ldots,$$

and for $z' \in E_Z^{i_z \wedge 2}$

$$T_{z'} f_{z'}(E_u^{i_u \wedge 1}) = f_{z'}(E_u^{i_u \wedge k+2}), \quad T_{z'} f_{z'}(E_u^{i_u \wedge 2}) = f_{z'}(E_u^{i_u \wedge 1}), \ldots$$

Then again by transfinite induction as in the previous case one obtains a Condorcet cycle $T_z(f_z(u))$ modulus sets of measure zero in the sense that $T_{z_i} f_{z_i}(u) \neq T_{z_i} f_{z_j}(u)$ almost all $u \in E_u$ and $z, z' \in E_Z$ as well as $z_1, \ldots, z_k$, as required.

Finally, the case where there are only finitely many atoms z_1, $z_2, \ldots, z_k$ such that $f_{z_j}(u) = f_{z_i}(u)$ for some Borel set $E_u \subset [0,1]$ with measure $P_U(E_u) = \varepsilon_u > 0$ is a special case of the above construction. The above construction covers all cases for uncountable or discrete subsets E_Z where $f_z(u)$ is not a one-to-one generator. In all cases we were able to construct a Condorcet cycle for all of $[0,1]$ by letting T_z be the identity map on all other sets except those sets E_Z; by definition, a Condorcet cycle is equivalent to a one-to-one generator. Therefore, we can apply Lemma 1 to finish the proof. □

B.3 Proof of Proposition 1

Proof. We focus on $P_{X|Z=z}$ as the case for $P_{Y|X^*=x}$ is completely analogous. We build a product measurable space $([0,1]^n, \bigotimes_n \mathscr{B}_{[0,1],n})$ and define the product measure $\bigotimes_{i=1}^n P_{X|Z=z_i}$ on it. Then it is not hard to see that the family $\{\bigotimes_{i=1}^n P_{X|Z=z_i} : n \in \mathbb{N}\}$ of finite dimensional product measures forms a projective family (Bauer 1996, Definition 35.2). Indeed, if $\{z_{n^*}\} \subset \{z_n\}$, then the product measure $\bigotimes_{i=1}^{n^*} P_{X|Z=z_i}$ is the marginal distribution of $\bigotimes_{i=1}^n P_{X|Z=z_i}$. We can therefore apply Kolmogorov's Extension Theorem (Bauer 1996, Theorem 35.3) to conclude that there exists a measure space $(\Omega, \mathcal{A}, P)$ and a family of random variables $X_z(\omega)$, $z \in [0,1]$ such that $P_{X|Z=z}$ is the law of the stochastic process $X_z(\omega)$. An analogous conclusion holds for $P_{Y|X^*=x}$ and $Y_x(\widetilde{\omega})$. □

B.4 Proof of Theorem 1

Proof. The construction in Hess (1982) is only for Wiener measure on $C([0,1])$. We therefore need to show that our measures $P_{Y|X^*}$ and $P_{X|Z}$ also induce stochastic processes which are defined on $C([0,1])$ under Assumption 3. In addition, we need to show that the linear program actually does what we want it to do, namely giving us bounds on the counterfactual probabilities. Let us start with the first issue. Showing this is similar to Hess (1982).

We focus on $P_{X|Z}$ and the corresponding construction since the case for $P_{Y|X^*}$ is perfectly analogous. Recall the following definitions from Lemmas 2 and 3 in the main text: $F_{n_1,\ldots,n_k}$ is a dyadic quasi-interval of order $k-1$, $D_{n_1,\ldots,n_k}$ is the disjoint version of these quasi-intervals from Lemma 2, $J \equiv J^X$ is the F_σ-set of closed subsets J_m of $\mathbb{I}$ from Lemma 3. Moreover, $B^m_{n_1,\ldots,n_k} \coloneqq F_{n_1,\ldots,n_k} \cap C^{0,\lambda}_m([0,1])$ and $\mathbb{I} \supset J_m \coloneqq \{(n_i)_{i\in\mathbb{N}} \in \mathbb{N}^{\mathbb{N}} : B^m_{n_1,\ldots,n_k} \neq \emptyset\}$.

We now show that if $P_{X|Z}$ satisfies Assumption 3, then there exists a measure ρ on $\mathbb{N}^{\mathbb{N}}$ with the following two properties: (i) $\rho(\langle n_1,\ldots,n_k\rangle) = P(F_{n_1,\ldots,n_k})$ for all $k \in \mathbb{N}$ and (ii) $\rho(J) = 1$, where P is the measure of the stochastic process X_z on the dyadic quasi-interval induced by $P_{X|Z}$ and $X = g(Z,U)$, $(n_1,\ldots,n_k) \in \mathbb{N}^k$. Part (i) is straightforward by setting $\rho(\langle n_1,\ldots,n_k\rangle) = P(F_{n_1,\ldots,n_k}) = P(D_{n_1,\ldots,n_k})$ (Hess 1982, p. 342). So let us show (ii). This can be done if we find a real valued function $q(m)$, $m \in \mathbb{N}$ such that $\rho(\mathbb{N}^{\mathbb{N}} \setminus J_m) \leq q(m)$ for all but finitely many m and $q(m) \to 0$ as $m \to \infty$. Hess (1982, p. 343) shows that

$$\mathbb{N}^{\mathbb{N}} \setminus J_m \subset \bigcup_{k=1}^{\infty} \left[\bigcup_{j=0}^{2^k-1} \bigcup_{D_{n_1,\ldots,n_k} \cap B^{c\cdot m}_{k,j} \neq \emptyset} \langle n_1,\ldots,n_k\rangle \right],$$

for $c \coloneqq \frac{1-2^{-\lambda}}{2}$ and where

$$B^h_{k,j} \coloneqq \left\{ f \in \mathbb{R}^{[0,1]} : \left| f_{\frac{j+1}{2^k}} - f_{\frac{j}{2^k}} \right| > h \cdot 2^{-k\lambda} \right\}.$$

This implies that

$$\begin{aligned}
\rho(\mathbb{N}^{\mathbb{N}} \setminus J_m) &\leq \sum_{k=1}^{\infty} \left[\sum_{j=0}^{2^k-1} \sum_{D_{n_1,\ldots,n_k} \cap B^{c\cdot m}_{k,j} \neq \emptyset} \rho(\langle n_1,\ldots,n_k\rangle) \right] \\
&= \sum_{k=1}^{\infty} \left[\sum_{j=0}^{2^k-1} \sum_{D_{n_1,\ldots,n_k} \cap B^{c\cdot m}_{k,j} \neq \emptyset} P(D_{n_1,\ldots,n_k}) \right] =: q(c\cdot m),
\end{aligned}$$

so that all we have to prove is $\lim_{m\to\infty} q(c\cdot m) = 0$.

But this follows from Assumption 3 and Chebychev's inequality just as in the proof of Kolmogorov's continuity theorem. In fact, we have for $\lambda \in (0, \frac{\delta}{\gamma})$ by Chebychev's inequality and

Assumption 3

$$\begin{aligned}
&\sum_{j=0}^{2^k-1} P\left(\left|X_{\frac{j+1}{2^k}} - X_{\frac{j}{2^k}}\right| > m \cdot c \cdot 2^{-k\lambda}\right) \\
\leq &\sum_{j=0}^{2^k-1} \frac{E\left(\left|X_{\frac{j+1}{2^k}} - X_{\frac{j}{2^k}}\right|^\gamma\right)}{m \cdot c \cdot 2^{-\lambda\gamma k}} \\
\leq &c_x \cdot m^{-1} \cdot c^{-1} \cdot 2^{-k(\delta - \lambda\gamma)}
\end{aligned}$$

But this implies that

$$\begin{aligned}
&\sum_{k=1}^{\infty}\left[\sum_{j=0}^{2^k-1} \sum_{D_{n_1,\ldots,n_k} \cap B_{k,j}^{c\cdot m} \neq \emptyset} P(D_{n_1,\ldots,n_k})\right] \\
\leq &\sum_{k=1}^{\infty}\left[\sum_{j=0}^{2^k-1} P\left(\left|X_{\frac{j+1}{2^k}} - X_{\frac{j}{2^k}}\right| > m \cdot c \cdot 2^{-k\lambda}\right)\right] \\
\leq &c_x \cdot m^{-1} \cdot c^{-1} \sum_{k=1}^{\infty} 2^{-k(\delta - \lambda\gamma)} < +\infty,
\end{aligned}$$

so that $q(c \cdot m) \to 0$ as $m \to \infty$. This proves that ρ satisfies (i) and (ii) and is hence an admissible measure on $\mathbb{N}^{\mathbb{N}}$. Finally, the thusly constructed measure is supported on $C([0,1])$ by an application of the Borel-Cantelli lemma in conjunction with the last inequality above, which implies that X_z possesses a modification with Hölder exponent $\lambda \in (0, \frac{\delta}{\gamma})$. In particular, let us denote $A_k \coloneqq \left\{\left|X_{\frac{j+1}{2^k}} - X_{\frac{j}{2^k}}\right| > m \cdot c \cdot 2^{-k\lambda}\right\}$. Then

$$P(A_k) \leq \sum_{j=0}^{2^k-1} P\left(\left|X_{\frac{j+1}{2^k}} - X_{\frac{j}{2^k}}\right| > m \cdot c \cdot 2^{-k\lambda}\right) < +\infty.$$

Now since $\sum_{k=1}^{\infty} P(A_k) < +\infty$, it holds by the Borel-Cantelli Lemma that $P\left(\bigcap_{l=1}^{\infty} \bigcup_{k=l}^{\infty} A_k\right) = 0$, i.e. this set is of probability zero. But this implies that there is a continuous modification such that

$$\left|X_{\frac{j+1}{2^k}} - X_{\frac{j}{2^k}}\right| \leq m \cdot c \cdot 2^{-k\lambda}$$

P-almost everywhere. Now the conclusion follows verbatim from the standard way to prove Kolmogorov's continuity theorem, for which we refer to Bauer (1996, Theorem 39.3).

Note that the measure is supported on every cylinder set, which follows from the equicontinuity of the sets $C_m^{0,\lambda}([0,1])$ and the fact that $\lim_{m\to\infty} P\left(C_m^{0,\lambda}([0,1])\right) = 1$, which in turn follow directly from our construction (Hess 1982, p. 346). The measure ρ constructed on $\mathbb{N}^{\mathbb{N}}$ can be regarded

as a nonatomic measure on $[0,1]$ such that $\rho(U) > 0$ for every non-empty open $U \subset [0,1]$, since we can identify the irrational numbers with Baire space. Furthermore, the Borel isomorphism ϕ_U from Lemma 3 is such that $\rho(E) = P(\phi_U^{-1}(E))$ for every Borel set $E \subset [0,1]$ (Hess 1982, p. 345). The analogous construction works for $P_{Y|X^*}$ and a measure ν analogous to ρ.

The measure μ from the statement of Theorem 1 must hence be such that its marginals $P_V = \pi_1\mu$ and $P_U = \pi_2\mu$ are such that $\phi_V^{-1}(P_V)$ and $\phi_X^{-1}(P_U)$ define proper probability measures on $C([0,1])$ which is just what $\mathcal{P}^*([0,1]^2)$ requires. We therefore define the measure μ as the induced measure of $(Y_{X_z(u)}(v), X_z(u))$, i.e. as $\mu_u \cdot \pi_2\mu$, where μ_u is the disintegrated measure defined for a fixed path $X_z(u)$. For fixed z, the path $X_z(u)$ gives a unique x, and μ_u is then induced by the map ϕ_V as before. The difference to $\pi_1\mu$ is that x is not externally determined, but determined by fixing a path $X_z(u)$—varying z gives different values for x, and based on those we define the measure induced by μ_u. In fact, this captures the property that the processes Y_x and X_z only depend on one another by the fact that Y_x depends on the position of X_z and hence on the value of u in general. This is hence a direct transformation of the nonseparable triangular model (1) to the two dependent stochastic processes Y_x and X_z. Now since ϕ_V and ϕ_U fix the unobservables v and u, the proof that (3), (5), and (6) are of the correct form can be accomplished in perfect analogy to the argumentation in Balke & Pearl (1994). For given values (y,x,z) the distribution $F_{Y,X|Z=z}(y,x)$ determines the probability that the continuous stochastic processes $Y_x(v)$ and $X_z(u)$ go through the sets $[0,y] \times [0,x]$ at z. $G_{Y,X}$ captures exactly this. Analogously, $H_{Y,X}$ captures all pairs of processes $Y_x(v)$ and $X_z(u)$ for which $Y_x(v)$ goes through $[0,y^*]$ at x_0. □

B.5 Proof of Proposition 3

Proof. We only need to prove that the constraints are equivalent. So consider the constraint from the problems (7)

$$f_{Y,X|Z=z}(y,x) = \int\int \Gamma(y,x,z,v,u)\mu(dv,du).$$

Now integrate both sides and apply Fubini's Theorem to obtain

$$\begin{aligned}
F_{Y,X|Z=z}(y,x) &= \int_{[0,y]\times[0,x]} f_{Y,X|Z=z}(s_y,s_x)ds_yds_x \\
&= \int_{[0,y]\times[0,x]} \int\int \Gamma(y,x,z,v,u)\mu(dv,du)ds_yds_x \\
&= \int_{[0,y]\times[0,x]} \int\int \delta\left(s_y - Y_{X_z(u)}(v)\right)\delta\left(s_x - X_z(u)\right)\mu(dv,du)ds_yds_x \\
&= \int\int \int_{[0,y]\times[0,x]} \delta\left(s_y - Y_{X_z(u)}(v)\right)\delta\left(s_x - X_z(u)\right)ds_yds_x\mu(dv,du) \\
&= \int\int \int_0^x \int \mathbb{1}_{[0,y]}(s_y)\delta\left(s_y - Y_{X_z(u)}(v)\right)\delta\left(s_x - X_z(u)\right)ds_yds_x\mu(dv,du)
\end{aligned}$$

$$
\begin{aligned}
&= \int\int\int \mathbb{1}_{[0,x]}(s_x)\mathbb{1}_{[0,y]}\left(Y_{X_z(u)}(v)\right)\delta\left(s_x - X_z(u)\right) ds_x \mu(dv, du) \\
&= \int\int \mathbb{1}_{[0,x]}\left(X_z(u)\right)\mathbb{1}_{[0,y]}\left(Y_{X_z(u)}(v)\right)\mu(dv, du) \\
&= \int\int G_{Y,X}(y, x, z, v, u)\mu(dv, du) \\
&\equiv \int_{[0,1]^2} G_{Y,X}(y, x, z, v, u)\mu(dv, du)
\end{aligned}
$$

where lines 6 and 7 follow from the property of the Dirac-delta distribution. Therefore, the CDF-constraint is the integrated PDF-constraint. From this, and the fact that a PDF uniquely defines a CDF and a CDF uniquely defines a PDF almost everywhere, we obtain the equivalency almost everywhere of the constraints and a fortiori the equivalency almost everywhere of the optimization problems (3) and (7). □

B.6 Proof of Lemma 4

Proof. Let Assumption 4 hold. In order for $\mathcal{A}(P_{Y,X|Z=z})$ to be non-empty we need to require that the latent stochastic processes Y_x and X_z satisfy Assumption 3, i.e.

$$
\begin{aligned}
&\int_{[0,1]} |Y_{x_1}(v) - Y_{x_2}(v)|^{\alpha}\, \pi_1\mu(dv) \leq c_y |x_1 - x_2|^{1+\beta} \quad \text{and} \\
&\int_{[0,1]} |X_{z_1}(u) - X_{z_2}(u)|^{\gamma}\, \pi_2\mu(du) \leq c_x |z_1 - z_2|^{1+\delta}
\end{aligned}
$$

for some fixed constants $c_y, c_x, \alpha, \beta, \gamma, \delta > 0$. We can write the constraint as

$$
\begin{aligned}
F_{Y,X|Z=z}(y, x) &= \int_{[0,1]^2} \mathbb{1}_{[0,y]\times[0,x]}\{s_y, s_x\} P_{Y,X|Z=z}(ds_y, ds_x) \\
&= \int_{[0,1]^2} \mathbb{1}_{[0,y]\times[0,x]}\left\{Y_{X_z(u)}(v), X_z(u)\right\}\mu(dv, du),
\end{aligned}
$$

that is, the constraint requirement is that $F_{Y,X|Z=z}$ be the pushforward law of the joint measure μ at $z \in [0, 1]$.

We now argue that every product measure $\mu \coloneqq \pi_1\mu \otimes \pi_2\mu$ with marginals satisfying Assumption 3 is in $\mathcal{A}(F_{Y,X|Z=z})$ for every $F_{Y,X|Z=z}$ satisfying Assumption 4. To see this, recall that by Assumption 4 it holds that

$$
\int_{[0,1]^2}\int_{[0,1]^2} |(s_y, s_x) - (t_y, t_x)|^{\eta_1} P_{Y,X|Z=z_1}(ds_y, ds_x) P_{Y,X|Z=z_2}(dt_y, dt_x) \leq c_{y,x}|z_1 - z_2|^{1+\eta_2}.
$$

We can then write

$$
\int_{[0,1]^2}\int_{[0,1]^2} |(s_y, s_x) - (t_y, t_x)|^{\eta_1} P_{Y,X|Z=z_1}(ds_y, ds_x) P_{Y,X|Z=z_2}(dt_y, dt_x)
$$

$$
\begin{aligned}
&= \int_{[0,1]^2} \left| \left(Y_{X_{z_1}(u)}(v), X_{z_1}(u) \right) - \left(Y_{X_{z_2}(u)}(v), X_{z_2}(u) \right) \right|^{\eta_1} \mu(dv, du) \\
&= \int_0^1 \int_0^1 \left| \left(Y_{X_{z_1}(u)}(v), X_{z_1}(u) \right) - \left(Y_{X_{z_2}(u)}(v), X_{z_2}(u) \right) \right|^{\eta_1} \mu_u(dv) \pi_2\mu(du) \\
&= \int_0^1 \left| Y_{X_{z_1}(u)}(v) - Y_{X_{z_2}(u)}(v) \right|^{\eta_1} \mu_u(dv) \int_0^1 |X_{z_1}(u) - X_{z_2}(u)|^{\eta_1} \, \pi_2\mu(du) \\
&= \int_0^1 \left| Y_{X_{z_1}(u)}(v) - Y_{X_{z_2}(u)}(v) \right|^{\eta_1} \pi_1\mu(dv) \int_0^1 |X_{z_1}(u) - X_{z_2}(u)|^{\eta_1} \, \pi_2\mu(du) \\
&= \int_0^1 |Y_{x_1}(v) - Y_{x_2}(v)|^{\eta_1} \, \pi_1\mu(dv) \int_0^1 |X_{z_1}(u) - X_{z_2}(u)|^{\eta_1} \, \pi_2\mu(du) \\
&\leq c_y |x_1 - x_2|^{1+\eta_3} c_x |z_1 - z_2|^{1+\eta_4} \\
&\leq c_y c_x |z_1 - z_2|^{1+\eta_4} \\
&\leq c_y c_x |z_1 - z_2|^{1+\eta_2} =: c_{y,x} |z_1 - z_2|^{1+\eta_2}.
\end{aligned}
$$

Here the second line follows from the fact that $F_{Y,X|Z=z}(y,x)$ is the induced joint law of the processes Y_x and X_z, as required by the constraint in the optimization problem. The third line follows from a disintegration of μ into μ_u and $\pi_2\mu$. The fifth line follows from our Assumption $\mu \coloneqq \pi_1\mu \otimes \pi_2\mu$, which implies that the disintegrated measure coincides with the marginal measure. The sixth line follows from the fact that $(V,U) \perp\!\!\!\perp Z$ by assumption and by our assumption that $\mu = \pi_1\mu \otimes \pi_2\mu$; because of this we can simply replace the expression $X_{z_1}(u)$ and $X_{z_2}(u)$ by some points x_1 and x_2 as the left integral does not depend on u. The seventh line follows by setting $\alpha = \gamma = \eta_1$ and $\beta = \delta = \eta_2$, and the eighth line follows from the fact that $|x_1 - x_2|^{1+\eta_2} \leq 1$ for all $x_1, x_2 \in [0,1]$ since $\eta_2 > 0$. Finally, the ninth line follows from $\eta_2 \leq \eta_4$ and the fact that $|z_1 - z_2| \leq 1$. □

B.7 Proof of Proposition 4

Proof. Let us first show convexity and compactness of $\mathcal{P}^*([0,1]^2)$ under Assumptions 2 and 3. As for convexity, recall from the proof of Theorem 1 that for given $\alpha, \beta, \gamma, \delta > 0$ and given $c_y, c_x > 0$ we can write Assumption 3 in terms of the measure μ from (3) as

$$
\begin{aligned}
&\pi_1\mu(\mathbb{N}^{\mathbb{N}} \setminus J_m^{\alpha,\beta}) \leq c_y \cdot m^{-1} \cdot C^{-1} \cdot \sum_{k=1}^{\infty} 2^{-k(\beta - \lambda_1\alpha)} < +\infty \quad \text{and} \\
&\pi_2\mu(\mathbb{N}^{\mathbb{N}} \setminus J_m^{\gamma,\delta}) \leq c_x \cdot m^{-1} \cdot C^{-1} \cdot \sum_{k=1}^{\infty} 2^{-k(\delta - \lambda_2\gamma)} < +\infty
\end{aligned}
\tag{18}
$$

for $\lambda_1 \in (0, \frac{\beta}{\alpha})$ and $\lambda_2 \in (0, \frac{\delta}{\gamma})$. For fixed $\alpha, \beta, \gamma, \delta, c_y, c_x > 0$, $\mathcal{P}^*([0,1]^2)$ is easily seen to be convex as the projection operators π_1 and π_2 are linear so that if $\mu_1, \mu_2 \in \mathcal{P}^*([0,1]^2)$, then $\pi_1\left((1-t)\mu_1 + t\mu_2\right) = (1-t)\pi_1\mu_1 + t\pi_1\mu_2$ and analogously for π_2. Therefore, for $(1-t)\mu_1 + t\mu_2$

we have

$$\pi_1\mu(\mathbb{N}^{\mathbb{N}} \setminus J_m^{\alpha,\beta}) = (1-t)\pi_1\mu_1(\mathbb{N}^{\mathbb{N}} \setminus J_m^{\alpha,\beta}) + t\pi_1\mu_2(\mathbb{N}^{\mathbb{N}} \setminus J_m^{\alpha,\beta}) \leq c_y \cdot m^{-1} \cdot C^{-1} \cdot \sum_{k=1}^{\infty} 2^{-k(\beta-\lambda_1\alpha)} < +\infty,$$

and analogously for π_2.

As for compactness, it follows from Prokhorov's theorem (e.g. Theorem 5.1 in Billingsley 1999) that the set of all probability measures on $[0,1]^2$, $\mathcal{P}([0,1]^2)$, is compact in the weak topology. We therefore only have to prove that $\mathcal{P}^*([0,1]^2)$ is closed in the weak topology to prove compactness. To prove closedness of $\mathcal{P}^*([0,1]^2)$, let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of probability measures in $\mathcal{P}^*([0,1]^2)$ converging weakly to some μ. Recall from Lemma 3 that the constructed $(1,1)$-homeomorphisms ϕ_V and ϕ_U are $(0,1)$-homeomorphisms from each $J_m^{\alpha,\beta}$ (respectively $J_m^{\gamma,\delta}$) onto the closed subsets $C_m^{\lambda}([0,1]) \subset C([0,1])$ for each m. Therefore, $\mathbb{N}^{\mathbb{N}} \setminus J_m^{\alpha,\beta}$ and $\mathbb{N}^{\mathbb{N}} \setminus J_m^{\gamma,\delta}$ are open for all m as $C([0,1]) \setminus C_m^{\lambda}([0,1])$ is open for all m. Now since μ_n converges weakly to μ, it follows from Portmanteau's Theorem that $\liminf_n \mu_n(O) \geq \mu(O)$ for all open Borel sets O (Billingsley 1999, Theorem 2.1), so that

$$\pi_1\mu(\mathbb{N}^{\mathbb{N}} \setminus J_m^{\alpha,\beta}) \leq \liminf_n \pi_1\mu_n(\mathbb{N}^{\mathbb{N}} \setminus J_m^{\alpha,\beta}) \leq c_y \cdot m^{-1} \cdot C^{-1} \cdot \sum_{k=1}^{\infty} 2^{-k(\beta-\lambda_1\alpha)}$$

$$\pi_2\mu(\mathbb{N}^{\mathbb{N}} \setminus J_m^{\gamma,\delta}) \leq \liminf_n \pi_2\mu_n(\mathbb{N}^{\mathbb{N}} \setminus J_m^{\gamma,\delta}) \leq c_x \cdot m^{-1} \cdot C^{-1} \cdot \sum_{k=1}^{\infty} 2^{-k(\delta-\lambda_2\gamma)},$$

so that $\mu \in \mathcal{P}^*([0,1]^2)$.

Let us now show non-emptiness and convexity of $\mathcal{A}(F_{Y,X|Z=z}(y,x))$ under Assumptions 2 – 4. The fact that $\mathcal{A}(F_{Y,X|Z=z}(y,x))$ is non-empty follows directly from Lemma 4. As for convexity of $\mathcal{E}([0,1]^2)$, let $\mu_1, \mu_2 \in \mathcal{E}([0,1]^2)$. Since the operator $\Theta\mu : \mathcal{M}([0,1]^2) \to C([0,1]^3)$ is linear it also holds that $\Theta[(1-t)\mu_1](y,x,z) + \Theta[t\mu_2](y,x,z) = F_{Y,X|Z=z}(y,x)$ for all $t \in (0,1)$. This, together with the convexity of $\mathcal{P}^*([0,1]^2)$ shows that $\mathcal{A}(F_{Y,X|Z=z}(y,x))$ is a convex-valued correspondence. Pre-compactness follows from the fact that it is a subset of the compact $\mathcal{P}^*([0,1]^2)$. □

B.8 Proof of Proposition 5

Proof. We focus on the minimization problem as the maximization is completely analogous. We prove that the value function $\underline{m}(\cdot)$ is convex, proper, and lower semicontinuous at $F_{Y,X|Z=z}$ from which strong duality follows. $\Xi\mu$ is convex and has a finite value for some $\mu \in \mathcal{P}^*([0,1]^2)$, which implies that $\underline{m}(F_{Y,X|Z=z}(y,x))$ is convex and proper if it is lower-semicontinuous by Theorem 4 in Rockafellar (1974). So all we need to show is lower semicontinuity. The function $\underline{m}(F_{Y,X|Z=z}(y,x))$ is lower semicontinuous if its epigraph $\text{epi}(\alpha) \coloneqq \{F_{Y,X|Z=z} \in C([0,1]^3) : \underline{m}(F_{Y,X|Z=z}(y,x)) \leq \alpha\}$, i.e. the set of all values greater or equal than the function value, is closed for any $\alpha \in \mathbb{R}$.

To show closedness of the epigraph of $\underline{m}(F_{Y,X|Z=z}(y,x))$ fix some $\alpha < +\infty$ first and let

$\{F_{Y,X|Z;n}\}_{n\in\mathbb{N}} \in \text{epi}(\alpha)$ be a sequence. Since $\{F_{Y,X|Z;n}\}_{n\in\mathbb{N}} \in \text{epi}(\alpha)$, it holds by definition of $\underline{m}(\cdot)$ that $F_{Y,X|Z;n}$ is a probability distribution function for all n, as general distribution functions cannot be replicated by a probability measure μ. Then by the uniform convergence it must be the case that $F_{Y,X|Z}$ is a probability distribution. Since $\underline{m}(F_{Y,X|Z=z;n}(y,x)) \leq \alpha$, it holds by definition of $\underline{m}$ that $F_{Y,X|Z=z;n}(y,x)$ is replicable by some $\mu_n \in \mathcal{P}^*([0,1]^2)$ in the sense that

$$F_{Y,X|Z=z;n}(y,x) = \int_{[0,1]^2} G_{Y,X}(y,x,z,v,u)\mu_n(dv,du).$$

We can therefore write

$$\begin{aligned} F_{Y,X|Z=z;n}(y,x) =& \int_{[0,1]^2} S_1(s_y,y)S_2(s_x,x)dF_{Y,X|Z=z;n}(s_y,s_x) + \delta \\ =& \int_{[0,1]^2} S_1\left(Y_{X_z(u)}(v),y,\varepsilon_1\right)\cdot S_2\left(X_z(u),x,\varepsilon_2\right)\mu_n(dv,du) + \delta \\ =& \mu_n\left(Y_x\in[0,y], X_z\in[0,x]\right) + \delta, \end{aligned} \tag{19}$$

where the approximation error $\delta \in (-\varepsilon,\varepsilon)$ for some small $\varepsilon > 0$ is due to the approximation of the indicator functions by logit functions.

The requirement (19) means that the measure $P_{Y,X|Z=z;n}$ associated with $F_{Y,X|Z=z;n}(y,x)$ is the pushforward of μ_n via (Y_x, X_z), so that different choices of $(Y_{X_z(u)}(v), X_z(u))$ yield different compatible μ_n. Furthermore, recall that $Y_x(v)$ and $X_z(u)$ are $(0,1)$-homeomorphisms for $C_c^{0,\lambda}([0,1])$ for some $c < +\infty$, which is fixed by Assumption 3, so that their Cartesian product $(Y_{X_z(u)}(v), X_z(u))$ is continuous. Since we assumed that $F_{Y,X|Z=z;n} \in \text{epi}(\alpha)$, it must be that $\mathcal{A}(F_{Y,X|Z=z;n}(y,x))$ is non-empty for all n by definition of $\underline{m}$. Now since $F_{Y,X|Z=z;n}(y,x)$ converges and is hence a Cauchy sequence, the requirement (19) implies that for every $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that for all $m,n \geq N$ there exist $\mu_m \in \mathcal{A}(F_{Y,X|Z=z;m}(y,x))$ and $\mu_n \in \mathcal{A}(F_{Y,X|Z=z;n}(y,x))$ with

$$\begin{aligned} &\sup_{(y,x,z)\in[0,1]^3} \left|\mu_m\left(Y_x\in[0,y], X_z\in[0,x]\right) - \mu_n\left(Y_x\in[0,y], X_z\in[0,x]\right)\right| \\ \leq& \sup_{(y,x,z)\in[0,1]^3} |F_{Y,X|Z=z;m}(y,x) - F_{Y,X|Z=z;n}(y,x)| + 2\delta \\ =& \|F_{Y,X|Z;m} - F_{Y,X|Z;n}\|_\infty + 2\delta < \varepsilon + 2\delta. \end{aligned}$$

But recall that the set of all rectangles induce the Borel σ-algebra, so that

$$\sup_{(y,x,z)\in[0,1]^3} \left|\mu_m\left(Y_x\in[0,y], X_z\in[0,x]\right) - \mu_n\left(Y_x\in[0,y], X_z\in[0,x]\right)\right| = \|\mu_m - \mu_n\|_{TV},$$

which, while letting $\delta \to 0$, implies that $\{\mu_n\}_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathcal{M}([0,1]^2)$ equipped with the topology induced by the total variation distance. Since $\mathcal{M}([0,1]^2)$ equipped with the

total variation norm is a Banach space, this sequence converges to some μ, which must satisfy (19). This implies that $\mathcal{A}(F_{Y,X|Z=z}(y,x))$ is non-empty. Moreover, since we did not restrict μ_n besides requiring (19), this holds for any sequence $\{\mu_n\}_{n\in\mathbb{N}} \subset \mathcal{P}^*([0,1]^2)$ satisfying (19).

Furthermore, recall the definition of Ξ

$$\int_{[0,1]^2} \Xi(y^*, x_0, v, u)\mu(dv, du) = \mu(Y_{x_0} \in [0, y^*]) + \delta = \pi_1\mu(Y_{x_0} \in [0, y^*]) + \delta,$$

where δ is again the approximation error. This implies, while letting $\delta \to 0$, by continuity of π_1 that $\|\pi_1\mu_m - \pi_1\mu\|_{TV} \to 0$. Now since the optimal value $\underline{m}(F_{Y,X|Z=z;n}(y,x))$ is not necessarily achieved by some μ_n^*, we can pick a μ_n' which is close enough in the sense that $\mu_n' \in \mathcal{A}(F_{Y,X|Z=z;n}(y,x))$ and $\Xi\mu_n' = \Xi\mu_n^* + \eta$ for some $0 < \eta < \varepsilon$ and every $n \in \mathbb{N}$. This is always possible since $S_1(Y_{X_z(u)}(v), y, \varepsilon_1)$ and $S_2(X_z(u), x, \varepsilon_2)$ are continuous in v and u, respectively and $\mathcal{P}^*([0,1]^2)$ is compact in the weak topology. In fact, since $\mathcal{A}(F_{Y,X|Z=z;N}(y,x))$ is weakly pre-compact, it must be the case that μ_N^* lies in the closure of $\mathcal{A}(F_{Y,X|Z=z;N}(y,x))$, so that there exists a sequence $\{\mu_n^m\}_{m\in\mathbb{N}}$ which lies in $\mathcal{A}(F_{Y,X|Z=z;N}(y,x))$, satisfies (19), and converges to μ_n^*. But since the above reasoning holds for any sequence of $\{\mu_n\}_{n\in\mathbb{N}}$, it must also hold for sequences $\{\mu_n'\}_{n\in\mathbb{N}}$, i.e. $\|\mu_n' \to \mu'\|_{TV} \to 0$ for some $\mu \in \mathcal{A}(F_{Y,X|Z=z}(y,x))$. Since we can make η as small as we want and by the continuity of π_1, this implies that $\underline{m}(F_{Y,X|Z=z;n}(y,x)) \to \underline{m}(F_{Y,X|Z=z}(y,x))$, which shows that $\underline{m}$ is lower semicontinuous.

To prove strong duality from this, let us write the Lagrangian $K(\mu,\nu)$ of the constrained optimization problem as

$$K(\mu,\nu) \coloneqq \Xi\mu + \langle F_{Y,X|Z} - \Theta\mu, \nu\rangle_2, \tag{20}$$

where $\nu \in \mathcal{M}([0,1]^3)$ is the Lagrange multiplier. Then we can write the problem in the following Lagrangian form

$$\inf_{\mu\in\mathcal{P}^*([0,1]^2)} \sup_{\nu\in\mathcal{M}([0,1]^3)} K(\mu,\nu). \tag{21}$$

The dual problem can then be written as

$$\sup_{\nu\in\mathcal{M}([0,1]^3)} \inf_{\mu\in\mathcal{P}^*([0,1]^2)} K(\mu,\nu). \tag{22}$$

Now note that for $f \in C([0,1]^3)$

$$\Phi(\mu, f) \coloneqq \begin{cases} \Xi\mu & \text{if} \quad \mu \in \mathcal{P}^*([0,1]^2) \quad \text{and} \quad F_{Y,X|Z=z}(y,x) - \Theta\mu - f = 0 \\ +\infty & \text{otherwise} \end{cases}$$

is convex in both its arguments, so that Theorem 7 in Rockafellar (1974) implies that

$$\sup_{\nu\in\mathcal{M}([0,1]^3)} \inf_{\mu\in\mathcal{P}^*([0,1]^2)} K(\mu,\nu) = \liminf_{f\to 0} \underline{m}(F_{Y,X|Z} + f).$$

But since $\underline{m}(\cdot)$ is lower semicontinuous it holds that $\liminf_{f\to 0} \underline{m}(F_{Y,X|Z}+f) = \underline{m}(F_{Y,X|Z})$, which is the value function of our constrained problem at $F_{Y,X|Z}$. This shows that there is no duality gap between the two problems. □

B.9 Proof of Proposition 6

Proof. For the proof we show directional Hadamard differentiability of $\underline{m}(\cdot)$ tangentially to $\mathcal{F}([0,1]^3)$. This, in combination with the Functional Delta Method and uniform weak convergence of the smoothed empirical processes, will then yield the result. As before, we give the proof for $\underline{m}(\cdot)$ as the proof for $\overline{m}(\cdot)$ is analogous.

First, note that under Assumptions 2 – 7 it follows from Proposition 4 in Giné & Nickl (2008) and the fact that the set of all rectangles in $[0,1]^3$ is a translation-invariant Donsker class that

$$\sqrt{n}\left(\hat{F}_{Y,X|Z;h_n} - F_{Y,X|Z}\right) \Rightarrow \mathbb{G},$$

where $\mathbb{G}$ is an $F_{Y,X|Z}$-Brownian bridge indexed by the set of all rectangles in $[0,1]^3$.

Second, we need to show that $\underline{m}(F_{Y,X|Z=z}(y,x))$ is continuous at every point of the set $\mathcal{F}([0,1]^3) \subset C([0,1]^3)$

$$\mathcal{F}([0,1]^3) \coloneqq \{F(y,x,z) \in C([0,1]^3) : F(y,x,z) \quad \text{satisfies Assumption 4}\}.$$

The proof for this is similar to the proof of Proposition 5. In fact, just as in this proof we use (19)

$$\begin{aligned} F_{Y,X|Z=z;n}(y,x) &= \int_{[0,1]^2} S_1(s_y,y)S_2(s_x,x)dF_{Y,X|Z=z;n}(s_y,s_x) + \delta \\ &= \int_{[0,1]^2} S_1\left(Y_{X_z(u)}(v),y,\varepsilon_1\right)\cdot S_2\left(X_z(u),x,\varepsilon_2\right)\mu_n(dv,du) + \delta \\ &= \mu_n\left(Y_x \in [0,y], X_z \in [0,x]\right) + \delta, \end{aligned}$$

for some approximation error $\delta \in (-\varepsilon,\varepsilon)$, $\varepsilon > 0$ which we set to zero for saving on notation. In Proposition 5 we already proved that $\underline{m}(\cdot)$ is lower semicontinuous on $C([0,1]^3)$. Now we prove that it is continuous on $\mathcal{F}([0,1]^3)$. We only need to show upper semicontinuity, i.e. we need to show that its hypograph $\text{hyp}(\alpha) \coloneqq \{F_{Y,X|Z=z} \in \mathcal{F}([0,1]^3) : \underline{m}(F_{Y,X|Z=z}(y,x)) \geq \alpha\}$ is closed.

To show this, note that by Lemma 4 $\mathcal{A}(F_{Y,X|Z=z}(y,x))$ is non-empty for every $F_{Y,X|Z=z}(y,x) \in \mathcal{F}([0,1]^3)$, so that $\underline{m}(F_{Y,X|Z=z}(y,x)) < +\infty$. Now let $\{F_{Y,X|Z;n}\}_{n\in\mathbb{N}} \subset \text{hyp}(\alpha)$ be some sequence converging to some $F_{Y,X|Z}$, that is $\|F_{Y,X|Z;n} - F_{Y,X|Z}\|_\infty \to 0$ as $n \to \infty$. Since $\mathcal{A}(F_{Y,X|Z;n}(y,x))$ is non-empty, there exist corresponding $\mu_n \in \mathcal{P}^*([0,1]^2)$ for which $F_{Y,X|Z;n}$ is the pushforward. Then by the same reasoning as in the proof of Proposition 5 this implies that $\|\mu_n - \mu\|_{TV} \to 0$ as $n \to \infty$ for some μ_n for which $F_{Y,X|Z=z;n}(y,x)$ are the pushforwards via some (Y_x, X_z). But since π_1 is a continuous operator, it must also hold that $\|\pi_1\mu_n - \pi_1\mu\|_\infty \to 0$ for these μ_n. Now, analogously

as in the proof of Proposition 5, we can set up a sequence $\{\mu_n^m\}_{m\in\mathbb{N}}$ for every $n \in \mathbb{N}$ which converges to the respectively optimal μ_n^* which might not be achieved. Then similar to before we can conclude that $\underline{m}(F_{Y,X|Z=z;n}(y,x)) \to \underline{m}(F_{Y,X|Z=z}(y,x))$, by continuity of the objective function. This shows upper semicontinuity of $\underline{m}(\cdot)$ on $\mathcal{F}([0,1]^3)$ and therefore continuity.

Let us now show Hadamard directional differentiability of $\underline{m}(\cdot)$ tangentially to $\mathcal{F}([0,1]^3)$. To do so, we need to show that the limit

$$\dot{\underline{m}}_{F_{Y,X|Z}}(F) = \lim_{n\to\infty} t_n^{-1}\left(\underline{m}(F_{Y,X|Z} + t_n F_n) - \underline{m}(F_{Y,X|Z})\right)$$

exists, where $\{t_n\}_{n\in\mathbb{N}} \in \mathbb{R}^+$ is a sequence of positive numbers converging to zero and $\{F_n\}_{n\in\mathbb{N}} \subset \mathcal{F}([0,1]^3)$ is a sequence converging to some $F \in \mathcal{F}([0,1]^3)$. Note that the subgradient of $\underline{m}(\cdot)$ on the contingent (Bouligand) cone[31] $\mathcal{T}_{F_{Y,X|Z}}(\mathcal{F}([0,1]^3))$ is defined as

$$\nu \in \partial\underline{m}(F_{Y,X|Z}) \quad \Leftrightarrow \quad \underline{m}(F_{Y,X|Z}+F) \geq \underline{m}(F_{Y,X|Z}) + t\langle F, \nu\rangle_2 \qquad \text{for all} \quad F \in \mathcal{F}([0,1]^3), \quad t > 0. \tag{23}$$

Based on this we now show that the Hadamard directional derivative tangentially to $\mathcal{F}([0,1]^3)$ exists and takes the form

$$\dot{\underline{m}}_{F_{Y,X|Z}}(F) = \max_{\nu\in\partial\underline{m}(F_{Y,X|Z})} \langle F, \nu\rangle_2 \qquad \text{for} \quad F \in \mathcal{F}([0,1]^3),$$

which is very similar to the conclusion in Theorem 11 of Rockafellar (1974), the only difference being that we require F to lie in the subset $\mathcal{F}([0,1]^3)$ and not the whole space $\mathcal{M}([0,1]^3)$. To do so first notice that (23) is equivalent to the statement

$$\nu \in \partial\underline{m}(F_{Y,X|Z}) \quad \Leftrightarrow \quad \dot{\underline{m}}_{F_{Y,X|Z}}(F) \geq \langle F, \nu\rangle_2 \qquad \text{for all} \quad F \in \mathcal{F}([0,1]^3),$$

as the limit $t_n \downarrow 0$ in the Hadamard directional derivative is equivalent to taking $\inf\{t > 0\}$ (Rockafellar 1974, p. 33). But then by definition it holds that

$$\dot{\underline{m}}_{F_{Y,X|Z}}(F) = \begin{cases} 0 & \text{if} \quad \nu \in \partial\underline{m}(F_{Y,X|Z}) \\ +\infty & \text{if} \quad \nu \notin \partial\underline{m}(F_{Y,X|Z}), \end{cases}$$

i.e. the Hadamard directional derivative is the indicator function from convex analysis of the subgradient $\partial\underline{m}_{F_{Y,X|Z}}(F)$, all of course contingent on $\mathcal{T}_{F_{Y,X|Z}}(\mathcal{F}([0,1]^3))$. The convex conjugate of an indicator function from convex analysis is the support function

$$\sup\{\langle F, \nu\rangle_2 : \nu \in \partial\underline{m}(F_{Y,X|Z})\},$$

[31] The contingent cone $\mathcal{T}_x(S)$ of a set $S \subset X$ in some Banach space X at a point $x \in X$ is defined as the set of all tangent elements $h \in X$ at x. An element h is tangent to S at x if there exists a sequence of elements $\{x_n\}_{n\in\mathbb{N}} \in X$ with $\lim_{n\to\infty} x_n = x$ and a sequence $\{r_n\}_{n\in\mathbb{N}} \in \mathbb{R}^+$ of positive elements such that $h = \lim_{n\to\infty} r_n(x - x_n)$.

see Rockafellar (1974, p. 16), so that the closure of the function $\dot{\underline{m}}_{F_{Y,X|Z}}(F)$ on $\mathcal{T}_{F_{Y,X|Z}}(\mathcal{F}([0,1]^3))$ coincides with the support function $\sup\{\langle F, \nu\rangle_2 : \nu \in \partial\underline{m}(F_{Y,X|Z})\}$, see Rockafellar (1974, p. 34). But since $\underline{m}(F)$ is continuous on $\mathcal{F}([0,1]^3)$, it holds that $\partial\underline{m}(F_{Y,X|Z})$ is weakly compact since its support function is continuous everywhere on $\mathcal{F}([0,1]^3)$ and hence equicontinuous (Rockafellar 1974, p. 34). Therefore, the supremum is attained and it holds that

$$\dot{\underline{m}}_{F_{Y,X|Z}}(F) = \max\{\langle F, \nu\rangle_2 : \nu \in \partial\underline{m}(F_{Y,X|Z})\}.$$

We can then apply the Functional Delta Method (Shapiro 1991, Theorem 2.1) to conclude that

$$\sqrt{n}\left(\underline{m}(\hat{F}_{Y,X|Z;n}) - \underline{m}(F_{Y,X,Z})\right) \Longrightarrow \dot{\underline{m}}_{F_{Y,X|Z}}(\mathbb{G}),$$

for $\hat{F}_{Y,X|Z;n} \in \mathcal{F}([0,1]^3)$ for all $n \in \mathbb{N}$. □

B.10 Proof of Proposition 7

Proof. We only focus on the minimization problem as the maximization is perfectly analogous. Note that y_j^* and $x_{0,j}$ in (15) are the closest smaller or equal dyadic points to y^* and x_0. Since the dyadic points are dense in $[0,1]$ it holds that $(y_j^*, x_{0,j}) \to (y^*, x_0)$ as $j \to \infty$. It is important that the dyadic points y_j^* and $x_{0,j}$ are *smaller* than y^* and x_0, since we make use of Dini's theorem below.

Now let $\{\mu_j\}_{j\in\mathbb{N}}$ be a sequence of feasible measures, i.e. measures satisfying the constraint (15). We first want to show that those measures converge weakly to a feasible measure μ_∞ satisfying

$$\left\| F_{Y,X|Z=z}(y,x) - \int \Theta(y,x,z,u,v)\mu_\infty(du,dv) \right\|^2_{L^2([0,1]^3)} \leq \varepsilon \tag{24}$$

for every $\varepsilon > 0$ as $j \to \infty$. Here $\|\cdot\|^2_{L^2([0,1]^3)}$ is the squared L^2-norm with respect to Lebesgue measure on $[0,1]^3$. We showed in the proof of Theorem 1 that for each dyadic decomposition j of the unit interval with m_j points, the marginal measures $\pi_1\mu_j(dv)$ and $\pi_2\mu_j(du)$ for a measure μ_j supported on those dyadic points induce measures $P_{Y|X\in\{x_i,\dots,x_{m_j}\}}$ and $P_{X|Z\in\{z_i,\dots,z_{m_j}\}}$. As the order j increases, these measures converge weakly to $P_{Y|X=x}$ and $P_{X|Z=z}$, respectively, for almost all x, z by the construction in Lemmas 2 and 3 and Theorem 1. Therefore by the Portmanteau theorem (Billingsley 1999, Theorem 2.1) and the mapping theorem (Billingsley 1999, Theorem 2.7) the measures μ_j converge weakly to μ_∞, because the constructed indices in Lemmas 2 and 3 are continuous by Assumption 3 for fixed c_y, c_x.

Furthermore, it holds that $\Theta_j(y_j, x_j, z_j, v, u) \to \Theta(y, x, z, v, u)$ for all v, u by Dini's theorem. In particular, recall that all Θ_j as well as Θ are a logistic approximation to the indicator function $\mathbb{1}_{[0,y]\times[0,x]}$. Now since y_j and x_j, and z_j are *smaller or equal* to y and x, and z by our construction of Θ_j, it holds that Θ_j converge monotonically to Θ in the sense that $\Theta_j \leq \Theta_{j+1}$. Moreover, the construction of the indices in Lemma 3 is a $(0,1)$-homeomorphism between $[0,1]$ and $C_c^{0,\lambda}([0,1])$

for some constant $c < +\infty$ only depending on λ from Assumption 3. Therefore, all Θ_j and Θ are continuous; in particular, they are uniformly continuous as $[0,1]$ is compact. Therefore, by Dini's theorem, it holds that $\Theta_j(y_j, x_j, z_j, u, v) \to \Theta(y, x, z, u, v)$ for all u, v. Hence, by weak convergence in conjunction with uniform convergence of the Θ_j it holds that

$$\int \Theta_j(y, x, z, v, u)\mu_j(dv, du) \to \int \Theta(y, x, z, v, u)\mu_\infty(dv, du),$$

where μ_∞ is the limit as $j \to \infty$. This last step is similar to part of the proof of Theorem 1 in Mendiondo & Stockbridge (1998).

$$F_{Y,X|Z=z}(y, x) - \int \Theta(y, x, z, v, u)\mu(dv, du)$$

is continuous for every $\mu \in \mathcal{P}^*([0,1]^2)$ by the construction of Θ and the assumption on $F_{Y,X|Z=z}$ and is therefore Riemann integrable. The dyadic points are asymptotically uniformly distributed on the unit interval. In particular, for any $\delta > 0$ there exists a large enough $j_0 \in \mathbb{N}$ with corresponding finite decomposition of $[0,1]$ into m_{j_0} subintervals $\{J_i, \ldots, J_{m_{j_0}}\}$ such that the upper and lower Darboux sums

$$\sum_{i=1}^{m_{j_0}} \frac{1}{m_{j_0}} t_i \text{Leb}(J_i) \qquad \text{and} \qquad \sum_{i=1}^{m_{j_0}} \frac{1}{m_{j_0}} s_i \text{Leb}(J_i)$$

are within δ of

$$\left\| F_{Y,X|Z=z}(y, x) - \int \Theta(y, x, z, u, v)\mu_\infty(du, dv) \right\|^2_{L^2([0,1]^3)}.$$

Here t_i (s_i) is the maximum (minimum) value of $F_{Y,X|Z=z}(y, x) - \int \Theta(y, x, z, v, u)\mu(dv, du)$ on the dyadic interval J_i.

Putting everything together, it therefore holds that for every $\eta > 0$ there exists $j_0 \in \mathbb{N}$ such that

$$\left| \frac{1}{m_{j_0}} \sum_{i=1}^{m_{j_0}} \left(F_{Y,X|Z=z_i}(y_i, x_i) - \int \Theta_{j_0}(y_i, x_i, z_i, u, v)\mu_{j_0}(du, dv) \right)^2 \right.$$
$$\left. - \left\| F_{Y,X|Z=z}(y, x) - \int \Theta(y, x, z, u, v)\mu_\infty(du, dv) \right\|^2_{L^2([0,1]^3)} \right| < \eta.$$

Recall that every μ_j satisfies the constraint of (15), and $\varepsilon_j \to 0$ as $j \to \infty$. Therefore, for every $\varepsilon > 0$ there is a large enough $j_0 \in \mathbb{N}$ such that $\eta < \varepsilon$. By assumption 4 we know that there exists a μ which satisfies the linear infinite dimensional constraint perfectly, so that (24) follows.

Now we need to show that a sequence $\{\mu_j^*\}$ of optimal measures converges to an optimal measure μ_∞^*. But this follows from the fact that Ξ_j are uniformly continuous and converge uniformly to Ξ by Dini's theorem. In particular, suppose by contradiction that μ_∞^* is not optimal,

i.e. there exists a feasible $\mu_0 \in \mathcal{P}^*([0,1]^2)$ such that

$$\int \Xi(y^*, x_0, z, v, u)\mu_0(dv, du) \leq \int \Xi(y^*, x_0, z, v, u)\mu^*_\infty(dv, du) - \delta$$

for some $\delta > 0$. Then note that we can project the measure μ_0 and its corresponding induced measures $P^0_{Y|X=x}$ and $P^0_{X|Z=z}$ onto the dyadic points for some order j. In particular, since those projections are continuous it follows that for any $\eta > 0$ there exists j_0 such that for all $j \geq j_0$

$$\left| \int \Xi_j(y^*_j, x_{0,j}, z_j, v, u)\mu_{0,j}(dv, du) - \int \Xi(y^*, x_0, z, v, u)\mu_0(dv, du) \right| < \eta$$

where $\mu_{0,j}$ denotes the projection of μ onto the dyadic points; this follows from the fact that all measures in $\mathcal{P}^*([0,1]^2)$, and so also μ_0, are tight, so that the finite projections onto the dyadic points are a determining class for μ_0 (Billingsley 1999, Example 1.3). But this implies directly that there must exist a large enough j_1 such that

$$\int \Xi_{j_1}(y^*_{j_1}, x_{0,j_1}, z_{j_1}, v, u)\mu_{0,j_1}(dv, du) < \int \Xi_{j_1}(y^*_{j_1}, x_{0,j_1}, z_{j_1}, v, u)\mu^*_{j_1}(dv, du)$$

if we choose $\eta < \delta$, which implies that μ^*_{j-1} is not optimal, a contradiction. □

B.11 Proof of Proposition 8

Proof. The proof of this proposition follows a similar reasoning to the proof of Theorem 2.1 in Pucci de Farias & Van Roy (2004), the difference being that we do not sample constraints like they do, but variables. We need to work with the dual programs of the linear programs in order to derive the result, as the sampled "variables" in the primal program correspond to sampled constraints in the dual program by standard duality arguments. Recall that the problems read

$$\begin{aligned}
&\underset{\mu \geq 0, \vec{1}'\mu \leq 1}{\text{minimize}} \quad \left(\widetilde{\Xi}^{min}_k\right)' \mu_{w,min} + \frac{\lambda_{min}}{2} \|\widetilde{\Theta}^{min}_k \mu_{w,min} - \hat{F}_{Y,X|Z=z;h_n}\|_2^2 \qquad \text{and} \\
&\underset{\mu \geq 0, \vec{1}'\mu \leq 1}{\text{maximize}} \quad \left(\widetilde{\Xi}^{max}_k\right)' \mu_{w,max} - \frac{\lambda_{max}}{2} \|\widetilde{\Theta}^{max}_k \mu_{w,max} - \hat{F}_{Y,X|Z=z;h_n}\|_2^2
\end{aligned}$$

Since we use the Euclidean norm for the penalization, these penalized programs can be written as

$$\begin{aligned}
&\min_{\mu \geq 0, \vec{1}'\mu \leq 1} \quad \frac{\lambda_{min}}{2} \mu' D^{min}_k \mu - \left(\lambda_{min} \left(\Theta^{min}_k\right)' \hat{F}_{Y,X|Z=z;h_n} - \Xi^{min}_k\right)' \mu + \frac{\lambda_{min}}{2} \left(\hat{F}_{Y,X|Z=z;h_n}\right)' \hat{F}_{Y,X|Z=z;h_n} \\
&\min_{\mu \geq 0, \vec{1}'\mu \leq 1} \quad \frac{\lambda_{max}}{2} \mu' D^{max}_k \mu - \left(\lambda_{max} \left(\Theta^{max}_k\right)' \hat{F}_{Y,X|Z=z;h_n} + \Xi^{max}_k\right)' \mu + \frac{\lambda_{max}}{2} \left(\hat{F}_{Y,X|Z=z;h_n}\right)' \hat{F}_{Y,X|Z=z;h_n},
\end{aligned}$$

where

$$D^{min}_k := \left(\Theta^{min}_k\right)' \Theta^{min}_k \qquad \text{and} \qquad D^{max}_k := \left(\Theta^{max}_k\right)' \Theta^{max}_k.$$

As stated, we need to consider the dual programs to the penalized programs since we do not want the vector μ to grow in dimension. These dual programs are quadratic functions in the dual variable y:

$$\min_{y\geq 0} \; -\frac{1}{2}y^2\vec{1}'\left(\lambda_{min}D_k^{min}\right)^{-1}\vec{1}+y\left[\vec{1}'\left(\lambda_{min}D_k^{min}\right)^{-1}\left[\lambda_{min}\left(\Theta_k^{min}\right)'\hat{F}_{Y,X|Z=z;h_n}-\Xi_k^{min}\right]-1\right]+R_{min}$$

$$\min_{y\geq 0} \; -\frac{1}{2}y^2\vec{1}'\left(\lambda_{max}D_k^{max}\right)^{-1}\vec{1}+y\left[\vec{1}'\left(\lambda_{max}D_k^{max}\right)^{-1}\left[\lambda_{max}\left(\Theta_k^{max}\right)'\hat{F}_{Y,X|Z=z;h_n}+\Xi_k^{max}\right]-1\right]+R_{max},$$

for

$$\begin{aligned}
R_{min} \coloneqq & \frac{\lambda_{min}}{2}\left(\hat{F}_{Y,X|Z=z;h_n}\right)'\hat{F}_{Y,X|Z=z;h_n} \\
& -\frac{1}{2}\left[\lambda_{min}\left(\Theta_k^{min}\right)'\hat{F}_{Y,X|Z=z;h_n}-\Xi_k^{min}\right]'\left(\lambda_{min}D_k^{min}\right)^{-1}\left[\lambda_{min}\left(\Theta_k^{min}\right)'\hat{F}_{Y,X|Z=z;h_n}-\Xi_k^{min}\right] \\
R_{max} \coloneqq & \frac{\lambda_{max}}{2}\left(\hat{F}_{Y,X|Z=z;h_n}\right)'\hat{F}_{Y,X|Z=z;h_n} \\
& -\frac{1}{2}\left[\lambda_{max}\left(\Theta_k^{max}\right)'\hat{F}_{Y,X|Z=z;h_n}+\Xi_k^{max}\right]'\left(\lambda_{max}D_k^{max}\right)^{-1}\left[\lambda_{max}\left(\Theta_k^{max}\right)'\hat{F}_{Y,X|Z=z;h_n}+\Xi_k^{max}\right].
\end{aligned}$$

Furthermore, since these programs are finite dimensional and well-behaved, strong duality holds, so that the optimal value of the primal program coincides with the optimal value of the dual program. The idea now is to exploit the finiteness of the VC-dimension (Vapnik & Chervonenkis 1971) of second-order polynomials on $\mathbb{R}$ in combination with Theorem 8.4.1 in Anthony & Biggs (1997). To be more specific, a standard result in complexity theory is that the VC-dimension of real polynomials of degree D in d variables is $\binom{d+D}{d}$, which follows from a bound on their shatter functions, see Theorem 5.5 in Matoušek (2009). In our case the degree of the dual programs is $D=2$ and the dimension is $d=1$.

The conclusion now follows from the same reasoning as in Pucci de Farias & Van Roy (2004), by applying Theorem 8.4.1 in Anthony & Biggs (1997). In particular, note that the hypothesis-set

$$\mathcal{H} \coloneqq \left\{\{(a_i,b_i,r_i)\in\mathbb{R}^3 : \frac{1}{2}a_iy^2+yb_i+r_i\leq 0\} : \; y\in\mathbb{R}\right\}$$

has VC-dimension 3 from what we have argued above. But now it follows directly from Theorem 8.4.1 in Anthony & Biggs (1997) that

$$\sup_{\{y:\, a_wy^2+yb_w+r_w\leq 0,\; w\in\mathcal{W}(s(\delta,\varepsilon))\}} P_s\left(\left\{i : a_iy^2+yb_i+r_i>0\right\}\right)\leq\varepsilon$$

if the set $\mathcal{W}(s(\delta,\varepsilon))$ has cardinality of at least $s(\delta,\varepsilon)$. Here the index i runs over all admissible paths for the given dyadic approximation of order j. Note that by changing the constant r

accordingly, we can change the hypothesis set to

$$\mathcal{H} \coloneqq \left\{ \{(a_i, b_i, r_i) \in \mathbb{R}^3 : \frac{1}{2} a_i y^2 + y b_i + r_i \leq V_k^{min}\} : \ y \in \mathbb{R} \right\}.$$

This implies that

$$\sup_{\{y:\, a_w y^2 + y b_w + r_w \leq V_k^{min},\, w \in \mathcal{W}_{s(\delta,\varepsilon)}\}} P_s \left(\left\{ i : a_i y^2 + y b_i + r_i > V_k^{min} \right\} \right) \leq \varepsilon,$$

so that the confidence of obtaining the optimal value with the sampled approach is $1 - \delta$. $\square$