Counterfactual Analysis for Structural Dynamic Discrete Choice Models[*]

Myrto Kalouptsidi, Yuichi Kitamura, Lucas Lima, and Eduardo Souza-Rodrigues[†]

September 2023

Abstract

Discrete choice data allow researchers to recover differences in utilities, but these differences may not suffice to identify policy-relevant counterfactuals of interest. In fact, in the case of dynamic discrete choice models, only a narrow set of counterfactuals are point-identified. In this paper, we explore how much one can learn about counterfactual outcomes of interest within this framework. We focus on the partial identification of counterfactuals, while allowing for (mild) model restrictions that can gradually shrink the identified set. We derive bounds for low-dimensional objects (such as average welfare) as arguments of optimization programs, along with a uniformly valid inference procedure. Furthermore, we develop new and tractable computational tools and algorithms suitable for dealing with high-dimensional problems like this. Finally, we illustrate in Monte Carlos, as well as an empirical exercise of firms' export decisions, the informativeness of the identified sets, and we assess the impact of (common) model restrictions on results.

KEYWORDS: Dynamic Discrete Choice, Counterfactual, Partial Identification, Structural Model.

[*] We would like to thank Victor Aguirregabiria, Phil Haile, Bo Honoré, Arthur Lewbel, Rob McMillan, Francesca Molinari, Ismael Mourifié, Salvador Navarro, Ariel Pakes, Aureo de Paula, John Rust, Elie Tamer, Xun Tang, Petra Todd, and Yuanyuan Wan for helpful discussions, and workshop participants at Cornell University, Ohio University, Saint Louis FED, UCLA, University of Toronto, the Gary Chamberlain Online Seminar in Econometrics, the ASSA Meetings 2019, the 3rd Conference on Structural Dynamic Models, the Counterfactuals with Economic Restrictions Conference at Western Ontario, the 5th Banff Empirical Microeconomics Workshop, the Interactions Conference: Bringing Together Econometrics and Applied Microeconomics at the University of Chicago, and the First IO Workshop at Javeriana University for additional comments. This paper circulated previously with the title "Partial Identification and Inference for Dynamic Models and Counterfactuals." Financial support from the University of Toronto Mississauga is gratefully acknowledged. All remaining errors are our own.

[†] Affiliations: Myrto Kalouptsidi, Harvard University, CEPR and NBER; Yuichi Kitamura, Yale University and Cowles Foundation for Research in Economics; Lucas Lima, Harvard University; Eduardo Souza-Rodrigues, University of Toronto.

1 Introduction

Discrete choice models have been used to answer a wide range of counterfactual questions in various fields of economics, including industrial organization, labor, public finance, and trade. It is well-known though that choice data allow researchers to recover only differences in individuals' valuations: in static models, we can identify differences in flow utilities (McFadden, 1974; Train, 2009); in dynamic models, we can recover differences in expected discounted streams of utilities (Rust, 1994; Magnac and Thesmar, 2002). In the latter case, which is the main focus of this paper, a recent literature has shown that knowing these differences in value functions does not suffice to identify many counterfactual outcomes of interest, as these may require knowledge of utility in levels; see Aguirregabiria and Suzuki (2014), Norets and Tang (2014), and Kalouptsidi, Scott, and Souza-Rodrigues (2021). The tension between identifying differences in valuations from choice data and the (potential) need for utility levels for counterfactuals can pose challenges to the credibility of structural estimation.

In this paper, we assert that partially identifying model parameters and counterfactuals is a natural route to proceed, and we explore how much one can learn about counterfactual outcomes of interest for a large (and empirically relevant) class of counterfactual experiments. At the same time, we allow for the possibility of incorporating additional (mild) assumptions on the underidentified payoff function that could gradually shrink the identified set. We also derive bounds for low-dimensional counterfactual objects, such as average welfare, as arguments of optimization programs, and provide an asymptotically uniformly valid inference procedure. We develop novel and tractable computational tools and algorithms capable of handling high-dimensional problems – a prevalent issue in applied studies. Our primary motivation is to offer a solution practitioners can use for a class of models used widely in empirical work.

To fix ideas, consider a typical example of a dynamic model encountered in applied work: a firm that decides every period whether to enter (exit) a market subject to entry costs (scrap values), with the goal of maximizing its lifetime payoffs consisting of variable profits minus fixed costs. Typically, researchers assume the payoff of staying out of the market (the 'outside option') is zero, and also impose that scrap values and/or fixed costs do not depend on state variables, and are equal to zero. While these assumptions suffice to identify flow payoffs, they may be strong for some industries and difficult to verify, since cost or scrap value data are extremely rare. Mistakenly setting the scrap value to zero, for example, lowers the firm's expected lifetime payoffs (since nothing is obtained each time the firm exits), which in turn generates a (possibly severe) downward bias in the estimated entry cost to rationalize the data. Most important, these assumptions are not always innocuous for important counterfactuals (Aguirregabiria and Suzuki, 2014; Norets and Tang, 2014; Kalouptsidi, Scott, and Souza-Rodrigues, 2021). Consider, for instance, a counterfactual exploring the impact of an entry cost subsidy: mistakenly setting the scrap value to zero not just potentially leads to a quantitatively wrong prediction of the subsidy's effect, but even to a wrong sign (see Kalouptsidi, Scott, and Souza-Rodrigues, 2021).

Our approach avoids such assumptions and bypasses estimating the model, focusing directly on the identified set of counterfactual objects (e.g., the welfare impact of a counterfactual entry subsidy) under much milder restrictions, such as positivity of the entry and fixed costs, and monotonicity of variable profits in demand shocks. In a numerical example (and in a Monte Carlo study), we illustrate that the identified sets are informative even under the mildest assumptions. We also explore which results survive under alternative model restrictions and show that common assumptions, such as zero scrap values, can be rejected.

We begin by showing that for a broad class of counterfactuals involving almost any change in the primitives, the sharp identified set for the counterfactual conditional choice probabilities (CCP) is a connected manifold with dimension that can be determined from the data, by checking the rank of a specific matrix, known to the econometrician. Specific combinations of model restrictions and counterfactual experiments can reduce the dimension of the identified set further, leading to point identification in some cases.

To aid practitioners, who may be interested in reducing the dimension of the identified set for both computational and policy-relevant reasons, we explore the reducing effect of some commonly used model restrictions. For instance, we explore parametric payoffs, counterfactuals that are "local" (i.e. that affect only a subset of the state and action space), as well as the combination of the two. We show that in all cases, the dimension of the identified set can be substantially reduced.

We then turn to low-dimensional outcomes of interest. We show that the sharp identified set here is also connected and, under additional mild conditions, compact. This is convenient as in practice it is sufficient to trace the boundary of the set. In addition, when the outcome of interest is a scalar, the identified set becomes a compact interval, in which case it suffices to calculate the lower and upper endpoints. The endpoints can be computed by solving well-behaved constrained minimization and maximization problems. The optimizations can be implemented using standard software (e.g., Knitro), and remain feasible even in high-dimensional cases involving large state spaces or a large number of model parameters.[1]

Our approach leads naturally to an inference procedure. We develop an asymptotically uniformly valid inference approach based on subsampling, and construct confidence sets based on test inversion. We also propose a novel computational algorithm for inference, which is tailored to our inherently high-dimensional dynamic setting. Such settings often present substantial challenges for other approaches in the partial identification literature (see, e.g., the discussion on computational challenges in Molinari, 2020). Yet, as demonstrated in the Monte Carlo study, our procedure is manageable even when the state space is large, and yields tight confidence sets with the correct coverage probabilities.

Overall, an attractive feature of this procedure is that the researcher can flexibly choose (i) the set of model restrictions, (ii) the counterfactual experiment, and (iii) the target outcome of interest, all without having to derive additional analytical identification results for each possible specification.

[1] For cases where computing the gradient is (prohibitively) costly, we develop and propose an alternative, stochastic search procedure that takes advantage of the structure of the problem (discussed in detail in the Appendix).

Finally, we illustrate the policy usefulness of our approach by revisiting Das, Roberts, and Tybout (2007), who perform a horserace between different types of export subsidies (export revenues, fixed cost, and entry cost subsidies). Similar to our firm entry/exit example, here firms decide whether to enter into/exit from exporting. Placing several restrictions on the model primitives (e.g., fixed costs and scrap values are equal to zero), they find that export revenue subsidies generate the highest net returns while entry cost subsidies result on the lowest returns. We obtain the identified sets for the net returns of the three subsidy measures under mild restrictions. We show that, although the ranking of Das, Roberts, and Tybout (2007) can be confirmed under weaker restrictions than originally imposed, it does hinge on the assumption that scrap values do not vary over states. Without this assumption, entry cost subsidies can potentially outperform the other types of subsidies.

Related Literature. A large body of work studies the identification and estimation of dynamic discrete choice (DDC) models. Rust (1994) showed that DDC models are not identified nonparametrically, and Magnac and Thesmar (2002) characterized the degree of underidentification. Estimation procedures were proposed by Rust (1987) and Hotz and Miller (1993).[2] We build on this literature on point-identification and estimation, and extend them to partial identification of model parameters and, more importantly, counterfactuals.

A small but growing literature investigates the identification of counterfactuals in DDC models; see Aguirregabiria (2010), Aguirregabiria and Suzuki (2014), Norets and Tang (2014), Arcidiacono and Miller (2020), and Kalouptsidi, Scott, and Souza-Rodrigues (2017, 2021).[3] Kalouptsidi, Scott, and Souza-Rodrigues (2021) (henceforth 'KSS') is our starting point. It provides the necessary and sufficient conditions for point identification of a broad class of counterfactuals and establishes that only a narrow set of counterfactuals is point-identified. Naturally, this result raises issues concerning the interpretation of empirical findings and poses an important challenge for researchers. Our paper provides a solution to this challenge, which relies on partially identifying counterfactuals under only mild assumptions. Furthermore, to the best of our knowledge, ours are the first analytical results characterizing the identified set of counterfactual behavior.

Aside from KSS, the closest paper to ours is by Norets and Tang (2014), who partially identify the structural parameters and the (high-dimensional) counterfactual CCPs in binary choice models with an unknown distribution of the idiosyncratic shocks. They focus on relaxing the distribution of the error term,

[2] These procedures were further analyzed by Hotz, Miller, Sanders, and Smith (1994); Aguirregabiria and Mira (2002, 2007); Bajari, Benkard, and Levin (2007); Pakes, Ostrovsky, and Berry (2007); and Pesendorfer and Schmidt-Dengler (2008). Important early contributions include Miller (1984), Wolpin (1984), and Pakes (1986). More recent related contributions by Dickstein and Morales (2018), Morales, Sheu, and Zahler (2019), and Berry and Compiani (2020) study partial identifiaction of model parameters in dynamic settings.

[3] Aguirregabiria and Suzuki (2014), Norets and Tang (2014), and Arcidiacono and Miller (2020) have established the identification of two important categories of counterfactuals in different classes of DDC models: counterfactual behavior is identified when flow payoffs change additively by pre-specified amounts; counterfactual behavior is generally not identified when the state transition process changes. Kalouptsidi, Scott, and Souza-Rodrigues (2017) discuss identification of counterfactual best-reply functions and equilibria in dynamic games.

while our focus is on the recovery of the counterfactual set of low-dimensional objects of interest, involving nonlinear functions of model parameters and counterfactual choice probabilities. Although nontrivial, our approach is computationally tractable, accommodating multinomial rather than just binary choice models. In terms of inference, Norets and Tang (2014) proposes a pointwise valid Bayesian approach, whereas our method relies on subsampling and is asymptotically uniformly valid. As such, our respective contributions are non-nested and complementary.[4]

Our inference approach builds on the formulation developed in Kitamura and Stoye (2018), where the implications of economic models are expressed in terms of the minimum value of a quadratic form. We consider a test statistic based on the minimum distance of the quadratic form to a kinked (i.e., non-regular), *random* (estimated), and possibly *nonconvex* set. We avoid standard convexity conditions on such objects because they are typically incompatible with our model restrictions.[5] We establish that an appropriate application of subsampling to the quadratic-form-based distance measure yields an asymptotically valid algorithm for inference.[6]

The inference procedure that we develop naturally relates to the work of Kaido, Molinari, and Stoye (2019) and Bugni, Canay, and Shi (2017), who provide general frameworks for uniformly valid inference procedures for low dimensional objects (e.g., a subvector or other functions of the model parameters) within moment inequality models. As such, they can be successfully applied to a wide range of empirically relevant problems. The nature of our proposal differs in that we tailor it to address challenges inherent to empirical studies utilizing dynamic discrete choice models. In particular, the main challenge here, which our approach is designed to overcome, is the high dimensionality of this setup – a prevalent issue in empirical applications. We elaborate further on this and explain the differences of these approaches in Section 6.

Taken together, we provide the first positive results on set-identification and computationally tractable valid inference procedures for counterfactual outcomes in structural dynamic multinomial choice models. These are the core contributions of our paper.

Finally, a recent and increasingly influential line of research emphasizes that (partial) identification of potential effects of policy interventions does not necessarily require identification of all the model parameters.[7] This line of research, including ours, is consistent with Marschak's (1953) prominent advocacy of

[4] Norets (2011) extends Norets and Tang (2014) to multinomial choice models, but the corresponding identified sets are infeasible to compute in practice. Our (set of) parametric distributions, in contrast, preserve the smoothness – and feasibility – of our constrained optimizations. Recently, and independently, Christensen and Connault (2021) proposed a clever way to perform sensitivity analysis of counterfactuals in structural models allowing for the distribution of unobservables to (nonparametrically) span neighborhoods of the researcher's assumed specification. When applied to DDC models, their approach complements ours, offering a promising avenue for future research.

[5] Note that Kitamura and Stoye (2018) deal with the case where a random vector is projected on a non-smooth but fixed object with some desirable geometric features. They then show that a bootstrap procedure combined with what they call the tightening technique leads to a computationally efficient algorithm with asymptotic uniform validity.

[6] Asymptotic validity of subsampling in nonregular models with more conventional settings, such as standard moment inequality models, have been shown in the literature: see Romano and Shaikh (2008) and Romano and Shaikh (2012).

[7] Contributions outside the class of structural dynamic models include Ichimura and Taber (2000, 2002) and Mogstad, Santos, and Torgovitsky (2018) for selection models; Manski (2007) for static choice models under counterfactual choice sets;

solving well-posed economic problems with minimal assumptions.

The rest of the paper is organized as follows: Section 2 sets out the framework; Section 3 contains our main results regarding the set-identification of counterfactual behavior, while Section 4 focuses on low-dimensional counterfactual outcomes of interest. Section 5 illustrates our results in the context of a firm entry/exit problem. Section 6 discusses estimation and inference, including a description of our computational algorithm. Section 7 presents the empirical application involving exporting subsidies; and Section 8 concludes.[8]

2 Framework

In this section we provide a brief description of the basic dynamic model and its identification, as well as define counterfactuals.

We assume time is discrete and the horizon is infinite. Every period t, agent i observes the state variables x_{it} and ε_{it}, and chooses an action $a \in \mathcal{A} = \{0, ..., A\}$, $A < \infty$, to solve

$$V\left(x_{it}, \varepsilon_{it}\right) = \max_{a \in \mathcal{A}} \left\{ \pi_a\left(x_{it}\right) + \varepsilon_{ait} + \beta\, \mathbb{E}\left[V\left(x_{it+1}, \varepsilon_{it+1}\right) | a, x_{it}, \varepsilon_{it}\right] \right\},$$

where $V(.)$ is the value function; $x_{it} \in \mathcal{X} = \{1, ..., X\}$, $X < \infty$, is observed by the econometrician and follows a controlled Markov process, $F(x_{it+1}|x_{it}, a)$; $\varepsilon_{it} = (\varepsilon_{0it}, ..., \varepsilon_{Ait})$ is not observed by the econometrician, is i.i.d. across agents and time, is independent of x_{it}, and has joint distribution G that is absolutely continuous with respect to the Lebesgue measure and has full support on $\mathbb{R}^{A+1}$; the per period utility from action a is additively separable in the bounded payoff function $\pi_a\left(x\right)$ and the unobservable ε_{ait}; and $\beta \in [0, 1)$ is the discount factor.

Following the literature, we define the *ex ante value function*, $V\left(x_{it}\right) \equiv \int V\left(x_{it}, \varepsilon_{it}\right) dG\left(\varepsilon_{it}\right)$, i.e. the expectation of $V\left(x_{it}, \varepsilon_{it}\right)$ over ε_{it}, as well as the *conditional value function*, $v_a\left(x_{it}\right) \equiv \pi_a\left(x_{it}\right) + \beta\, \mathbb{E}\left[V\left(x_{it+1}\right) | a, x_{it}\right]$. The *conditional choice probability* (CCP) function is given by:

$$p_a\left(x_{it}\right) = \int 1\left\{v_a\left(x_{it}\right) + \varepsilon_{ait} \geq v_j\left(x_{it}\right) + \varepsilon_{jit}, \text{ for all } j \in \mathcal{A}\right\} dG\left(\varepsilon_{it}\right),$$

Blundell, Browning, and Crawford (2008), Blundell, Kristensen, and Matzkin (2014), Kitamura and Stoye (2019), and Adams (2020) for bounds on counterfactual demand distributions and welfare analysis; Adao, Costinot, and Donaldson (2017) for international trade models; and Bejara (2020) for macroeconomic models.

[8] The Online and the Supplemental Material complement the main paper. The Online Appendix contains (a) all proofs of the propositions and theorems presented in the main text; (b) our proposed stochastic search approach to calculate the lower and upper bounds of the identified set of relevant outcomes, without analytic gradients; (c) the details of the computational algorithm for inference based on subsampling; and (d) our Monte Carlo study. The Supplemental Material presents (e) several useful examples of commonly employed restrictions in applied work (using our notation); (f) detailed information about our running example (the firm entry/exit problem); (g) the extension of our results to the case where neither the discount factor nor the distribution of the error term is known by the econometrician; (h) the analytical gradient of the counterfactual object of interest function when it involves long-run average effects; and (i) our replication of Das, Roberts, and Tybout (2007). The Supplemental Material is available on the authors' webpages.

where $1\{\cdot\}$ is the indicator function. We define the $(A+1) \times 1$ vector of CCPs $p(x) = (p_0(x), ..., p_A(x))'$, and the corresponding $(A+1)X \times 1$ vector $p = (p'(1), ..., p'(X))'$, where $'$ denotes transpose.

It is useful to note that for any (a, x) there exists a real-valued function $\psi_a(.)$ derived only from G such that

$$V(x) = v_a(x) + \psi_a\big(p(x)\big). \tag{1}$$

Equation (1) states that the ex ante value function V equals the conditional value function of any action a, v_a, plus a correction term, ψ_a, because choosing action a today is not necessarily optimal. When ε_{it} follows the type I extreme value distribution, we have that $\psi_a(p(x)) = \kappa - \ln p_a(x)$, where κ is the Euler constant. Chiong, Galichon, and Shum (2016) propose a computationally tractable approach, based on linear-programming, that can calculate ψ_a for any given distribution G. (See also Dearing 2019.)[9]

As we make extensive use of matrix notation below, we define the vectors $\pi_a, v_a, V, \psi_a \in \mathbb{R}^X$, which stack $\pi_a(x)$, $v_a(x)$, $V(x)$, and $\psi_a(p(x))$, for all $x \in \mathcal{X}$. We often use the notation $\psi_a(p)$ to emphasize the dependence of ψ_a on the choice probabilities p. We also define F_a as the transition matrix with (m, n) element equal to $\Pr\big(x_{it+1} = x_n | x_{it} = x_m, a\big)$. The payoff vector $\pi \in \mathbb{R}^{(A+1)X}$ stacks π_a for all $a \in \mathcal{A}$, and, similarly, F stacks (a vectorized version of) F_a for all $a \in \mathcal{A}$. We collect all primitives of the model in the tuple $s = (\mathcal{A}, \mathcal{X}, \beta, G, F, \pi)$.

Useful Representation. Following KSS, we have that, for all $a \neq J$, where $J \in \mathcal{A}$ is some (arbitrary) reference action, π_a can be represented as an affine transformation of π_J:[10]

$$\pi_a = M_a(F)\,\pi_J + b_a(p, F), \tag{2}$$

where

$$\begin{aligned} M_a(F) &= (I - \beta F_a)(I - \beta F_J)^{-1}, && (3) \\ b_a(p, F) &= M_a(F)\,\psi_J(p) - \psi_a(p), && (4) \end{aligned}$$

and I is a (comformable) identity matrix. In the logit model, $b_a(p, F) = \ln p_a - M_a(F) \ln p_J$, where $\ln p_a$ is the $X \times 1$ vector with elements $\ln p_a(x)$. To simplify notation, we omit the dependence of both matrix M_a and vector b_a on the discount factor β, as well as the dependence of b_a and ψ_a on G. We also omit

[9] Equation (1) is shown in Arcidiacono and Miller (2011, Lemma 1). It makes use of the Hotz-Miller inversion (Hotz and Miller, 1993), which, in turn, establishes that the difference of conditional value functions is a known function of the CCPs: $v_a(x) - v_j(x) = \varphi_{aj}\big(p(x)\big)$, where $\varphi_{aj}(.)$ is again derived only from G. When ε_{it} follows the type I extreme value distribution, $\varphi_{aj}(p(x)) = \log p_a(x) - \log p_j(x)$.

[10] To see why, fix the vector $\pi_J \in \mathbb{R}^X$. Then,

$$\pi_a = v_a - \beta F_a V = V - \psi_a - \beta F_a V = (I - \beta F_a) V - \psi_a,$$

where for $a = J$, we have $V = (I - \beta F_J)^{-1}(\pi_J + \psi_J)$. After substituting for V, we obtain the result. As an aside, note that $(I - \beta F_J)$ is invertible because F_J is a stochastic matrix and hence its largest eigenvalue is smaller than or equal to one.

their dependence on the transition probabilities F when it is sufficiently clear from the context.

Example. Suppose an agent faces a binary choice $a \in \{0,1\}$ and that ε_{it} follows the type I extreme value distribution. Equation (2) then becomes

$$\pi_1 = M_1 \pi_0 + \underbrace{(\log p_1 - M_1 \log p_0)}_{=b_1(p)}, \tag{5}$$

where $M_1 = (I - \beta F_1)(I - \beta F_0)^{-1}$.

It is instructive to compare this to a static model: in that case, since $\beta = 0$ (or, alternatively, $F_0 = F_1$ when choices do not affect future states), we have $M_1 = I$ and (2) becomes,

$$\pi_1 = \pi_0 + \underbrace{(\log p_1 - \log p_0)}_{=b_1(p)}. \tag{6}$$

In words, we obtain the well-known result that the difference of payoffs $\pi_1 - \pi_0$ equals the log odds ratio of the choice probabilities.

The difference between the dynamic model (5) and the static model (6) is the matrix M_1, which depends on the discount factor and the state transitions. This matrix distorts both the payoff difference and the log odds ratio of the CCPs (compared to the static version) to capture the impact of agents' expectations about the future.

By stacking (2) for all $a \neq J$ and rearranging, we obtain the useful compact representation:

$$\mathbf{M}(F)\,\pi = b_{-J}(p, F), \tag{7}$$

where $\mathbf{M}(F) = [I, -M_{-J}(F)]$; $M_{-J}(F)$ stacks $M_a(F)$ for all $a \neq J$; the payoff vector is arranged as $\pi = [\pi'_{-J}, \pi'_J]'$, where π_{-J} stacks π_a for all $a \neq J$; and the vector $b_{-J}(p, F)$ likewise stacks $b_a(p, F)$ for all $a \neq J$.[11] Equation (7) can be understood as an alternative representation to the Bellman equation, relating the primitives of the model and the conditional choice probabilities directly.[12]

Model Restrictions. We consider two types of model restrictions, beyond the basic setup. The first is a set of $d \leq X$ linearly independent equalities,

$$R^{eq}\pi = r^{eq}, \tag{8}$$

[11] This model imposes a scale normalization. In general, the flow utility is given by $\pi_a(x_{it}) + \sigma\varepsilon_{ait}$, where $\sigma > 0$ is a scale parameter. This means equation (7) becomes $\mathbf{M}\,(\pi/\sigma) = b_{-J}$. As usual in discrete choice models, when we set $\sigma = 1$ (as we do here), the scale of the payoff is measured relative to the standard deviation of one of the components of ε_{it}.

[12] To see why, note that the CCP vector p generated by the model primitives is the unique vector that satisfies (7): since the Bellman is a contraction mapping, V is unique; and hence so are v_a and p.

with $R^{eq} \in \mathbb{R}^{d\times(A+1)X}$, or in block-form, $R^{eq} = [R^{eq}_{-J}, R^{eq}_{J}]$, where R^{eq}_{-J} defines how π_{-J} enters into the constraints and, similarly, R^{eq}_{J} for π_J. This formulation is general enough to incorporate several assumptions used in practice. Examples include exclusion restrictions (setting some elements of π equal to each other), prespecifying some π_J (set $R^{eq}_{J} = I$, $R^{eq}_{-J} = 0$ and r^{eq} accordingly), and parametric assumptions.

The second set of restrictions are m linear inequalities:

$$R^{iq}\pi \leq r^{iq}, \tag{9}$$

with $R^{iq} \in \mathbb{R}^{m\times(A+1)X}$, or in block-form, $R^{iq} = [R^{iq}_{-J}, R^{iq}_{J}]$. The inequalities (9) can incorporate shape restrictions, such as monotonicity, concavity, and supermodularity. In Appendix E, we explicitly lay out how several examples used in applied work can be expressed as (8) or (9).

We assume (8) and (9) are not redundant. Therefore, equations (7), (8), and (9) summarize all the model restrictions.

Model Identification. Typically, the researcher has access to panel data on agents' actions and states, $\{a_{it}, x_{it} : i = 1, ..., N; t = 1, ..., T\}$. Under some standard regularity conditions, the researcher can identify and estimate the agents' choice probabilities p, as well as the transition function F, directly from the data. We therefore take p and F as known for the identification arguments. We also follow the literature and assume for now that the econometrician knows the discount factor β and the distribution of the idiosyncratic shocks G; we relax these assumptions in Appendix G. Under these conditions, the only remaining primitive to be identified from the data is the payoff function π.

The model is identified if there is a unique payoff that can be inferred from the observed choice probabilities and state transitions. From (7), it is clear that the basic dynamic setup alone would suffice to identify π were the matrix $\mathbf{M}$ invertible. However, identification fails because $\mathbf{M}$ is rank-deficient: $\mathbf{M}$ is an $AX \times (A+1)X$ matrix, and so $rank(\mathbf{M}) = AX < (A+1)\,X$. Intuitively, π has $(A+1)X$ parameters, and there are only AX observed CCPs; thus there are X free payoff parameters and X restrictions will need to be imposed to point-identify π (Rust, 1994; Magnac and Thesmar, 2002).

The sharp identified set for the payoff function is therefore the convex polyhedron characterized by all payoffs satisfying all model restrictions. Specifically, for $(p, F) \in \mathbf{P} \times \mathbf{F}$, where $\mathbf{P}$ is the simplex of conditional choice probabilities and $\mathbf{F}$ is the set of controlled Markovian transitions, the identified set for π is given by

$$\Pi^I(p, F) = \left\{ \pi \in \mathbb{R}^{(A+1)X} : \mathbf{M}(F)\,\pi = b_{-J}(p, F),\ R^{eq}\pi = r^{eq},\ R^{iq}\pi \leq r^{iq} \right\}. \tag{10}$$

This set has dimension $X - d$, where $0 \leq d \leq X$; point-identification is obtained when $d = X$ model restrictions are imposed.[13]

Counterfactuals. Next, we turn to counterfactuals. A counterfactual is defined by a transformation of the primitives. Formally, the counterfactual structure $\widetilde{s} = (\widetilde{\mathcal{A}}, \widetilde{\mathcal{X}}, \widetilde{\beta}, \widetilde{G}, \widetilde{F}, \widetilde{\pi})$ is obtained by applying the transformation $h = (h_{\mathcal{A}}, h_{\mathcal{X}}, h_{\beta}, h_G, h_F, h_{\pi})$ to the primitives $s = (\mathcal{A}, \mathcal{X}, \beta, G, F, \pi)$; i.e., $\widetilde{s} = h(s)$. The sets $\widetilde{\mathcal{A}} = h_{\mathcal{A}}(\mathcal{A}) = \{0, ..., \widetilde{A}\}$ and $\widetilde{\mathcal{X}} = h_{\mathcal{X}}(\mathcal{X}) = \{1, ..., \widetilde{X}\}$ denote the new set of actions and states, respectively. The new discount factor is $\widetilde{\beta} = h_{\beta}(\beta)$, the new distribution of the idiosyncratic shocks is $\widetilde{G} = h_G(G)$, and the new transition probability is $\widetilde{F} = h_F(F)$. Finally, the function $h_{\pi} : \mathbb{R}^{(A+1)X} \to \mathbb{R}^{(\widetilde{A}+1)\widetilde{X}}$ transforms the payoff function π into the counterfactual payoff $\widetilde{\pi}$. Here, we restrict transformations on payoffs to affine changes

$$\widetilde{\pi} = \mathcal{H}\pi + g, \tag{11}$$

where the matrix $\mathcal{H}$ and the vector g are specified by the econometrician. I.e., the payoff $\widetilde{\pi}_a(x)$ at an action-state pair (a, x) is obtained as the sum of a scalar $g_a(x)$ and a linear combination of all baseline payoffs.[14] The new set of model primitives $\widetilde{s}$ leads to a new lifetime utility, denoted by $\widetilde{V}$, and a new optimal behavior, denoted by $\widetilde{p}$. In practice, most applied papers consider counterfactuals that affect one primitive; see KSS for several examples of empirical work implementing such counterfactuals.

3 Identification of Counterfactual Behavior

We now investigate the identified set for the counterfactual CCP. To do so, we leverage the counterfactual counterpart to (2) for any action $a \in \widetilde{\mathcal{A}}$, with $a \neq J$. I.e.,

$$\widetilde{\pi}_a = \widetilde{M}_a(\widetilde{F})\, \widetilde{\pi}_J + \widetilde{b}_a(\widetilde{p}, \widetilde{F}), \tag{12}$$

where $\widetilde{M}_a(\widetilde{F}) = (I - \widetilde{\beta}\widetilde{F}_a)\,(I - \widetilde{\beta}\widetilde{F}_J)^{-1}$; $\widetilde{b}_a(\widetilde{p}, \widetilde{F}) = \widetilde{M}_a(\widetilde{F})\, \widetilde{\psi}_J\,(\widetilde{p}) - \widetilde{\psi}_a\,(\widetilde{p})$; the functions $\widetilde{\psi}_J$ and $\widetilde{\psi}_a$ depend on the new distribution $\widetilde{G}$ (which is omitted in our notation); and, without loss of generality, the reference action J belongs to both $\mathcal{A}$ and $\widetilde{\mathcal{A}}$. As before, we omit the dependence of both $\widetilde{M}_a$ and $\widetilde{b}_a$ on the discount factor $\widetilde{\beta}$ to simplify notation.

By stacking equation (12) for all actions in $\widetilde{\mathcal{A}}$, rearranging it as we did previously for the baseline case, and utilizing the fact that $\widetilde{\pi} = \mathcal{H}\pi + g$ and $\widetilde{F} = h_F(F)$, we obtain our main counterfactual equation:

$$(\widetilde{\mathbf{M}}(F)\mathcal{H})\, \pi = \widetilde{b}_{-J}(\widetilde{p}, F) - \widetilde{\mathbf{M}}(F)g, \tag{13}$$

[13] In the presence of unobserved heterogeneity, equations (7)–(9) hold for each unobserved type. This implies that, after type-specific choice probabilities and transition functions of finitely many unobserved types are identified (e.g., following the strategies proposed by Kasahara and Shimotsu (2009) or Hu and Shum (2012)), identified sets given by Π^I hold, and can be calculated, for each type.

[14] Extensions to nonlinear transformations of π are straightforward but not pursued here because they are not common in the empirical literature.

where $\widetilde{\mathbf{M}}(F) = \widetilde{\mathbf{M}}(h_F(F)) = [I, -\widetilde{M}_{-J}(h_F(F))]$; $\widetilde{M}_{-J}(h_F(F))$ stacks $\widetilde{M}_a(h_F(F))$ for all $a \neq J$; and $\widetilde{b}_{-J}(\widetilde{p}, F) = \widetilde{b}_{-J}(\widetilde{p}, h_F(F))$, where the vector $\widetilde{b}_{-J}(p, h_F(F))$ likewise stacks $\widetilde{b}_a(p, h_F(F))$ for all $a \neq J$.

Similar to (7), equation (13) is useful because it characterizes counterfactual behavior, relating (the unique) $\widetilde{p}$ and model parameters directly, with no continuation values involved. Importantly, $\widetilde{b}_{-J}$ is a continuously differentiable function of $\widetilde{p}$ (holding F fixed) with an everywhere invertible Jacobian (see Lemma 1 in KSS).

Let $\widetilde{\mathbf{P}}$ be the conditional probability simplex in $(\widetilde{A}+1)\widetilde{X}$. Our first proposition follows (all proofs are in the Appendix).

Proposition 1. *The sharp identified set for the counterfactual CCP $\widetilde{p}$ is*

$$\widetilde{\mathbf{P}}^I(p, F) = \left\{ \begin{array}{c} \widetilde{p} \in \widetilde{\mathbf{P}} : \exists \pi \in \mathbb{R}^{(A+1)X} \text{ such that} \\ \mathbf{M}(F)\,\pi = b_{-J}(p, F), \\ R^{eq}\pi = r^{eq}, \;\; R^{iq}\pi \leq r^{iq}, \\ (\widetilde{\mathbf{M}}(F)\mathcal{H})\,\pi = \widetilde{b}_{-J}\,(\widetilde{p}, F) - \widetilde{\mathbf{M}}(F)g \end{array} \right\}, \tag{14}$$

for $(p, F) \in \mathbf{P} \times \mathbf{F}$. The set $\widetilde{\mathbf{P}}^I(p, F)$ is a smooth connected manifold with boundary, and dimension in the interior given by the rank of the matrix $\mathcal{C}_J\mathcal{Q}_J$, where

$$\mathcal{C}_J = \underbrace{\left[I, -\widetilde{M}_{-J}(F)\right]}_{=\widetilde{\mathbf{M}}(F)} \mathcal{H} \begin{bmatrix} M_{-J}(F) \\ I \end{bmatrix}, \tag{15}$$

and $\mathcal{Q}_J$ is a known matrix (defined in the proof; see equation (A4) *in the Appendix) that depends on the model restrictions and on the transition probabilities F. Furthermore, rank$(\mathcal{C}_J\mathcal{Q}_J) \leq X - d$. In the absence of equality restrictions (8), the dimension of $\widetilde{\mathbf{P}}^I$ is given by rank$(\mathcal{C}_J) \leq X$. The inequality restrictions* (9) *do not affect the dimension of $\widetilde{\mathbf{P}}^I$.*[15]

In words, a vector $\widetilde{p}$ lying in the conditional probability simplex $\widetilde{\mathbf{P}}$ belongs to the identified set $\widetilde{\mathbf{P}}^I(p, F)$ if there exists a payoff π that is compatible with the data (i.e., $\mathbf{M}(F)\,\pi = b_{-J}(p, F)$), satisfies the additional model restrictions (i.e., $R^{eq}\pi = r^{eq}$ and $R^{iq}\pi \leq r^{iq}$), and can generate $\widetilde{p}$ in the counterfactual scenario (i.e., $(\widetilde{\mathbf{M}}(F)\mathcal{H})\,\pi = \widetilde{b}_{-J}\,(\widetilde{p}, F) - \widetilde{\mathbf{M}}(F)g$).

Intuitively, equation (13) implicitly defines $\widetilde{p}$ as a continuously differentiable function of π (taking the other primitives as given). The sharp identified set $\widetilde{\mathbf{P}}^I$ is therefore the image of Π^I under this function. It is clear that $\widetilde{\mathbf{P}}^I$ is empty whenever Π^I is empty (i.e., whenever the model is rejected in the data). An implication of the connectedness of the identified set is that a non-empty $\widetilde{\mathbf{P}}^I$ is either a singleton (in which case $\widetilde{p}$ is point-identified) or a continuum.

[15] Note that the indexing of the matrices $\mathcal{C}_J$ and $\mathcal{Q}_J$ by J does not affect their rank. The choice of the reference action only determines the arrangement of these matrices and is therefore arbitrary. This is an important point to keep in mind, as it emphasizes that the model's structure and identification are not dependent on this choice.

Proposition 1 also determines that the dimension of the identified set is always smaller than $X - d$, which is usually smaller than the dimension of $\widetilde{\mathbf{P}}$, given by $\widetilde{X}\widetilde{A}$. This implies that, typically, the identified set $\widetilde{\mathbf{P}}^I$ is informative. When $rank(\mathcal{C}_J\mathcal{Q}_J) = 0$, $\widetilde{\mathbf{P}}^I$ collapses to a singleton, which means that all points $\pi \in \Pi^I$ map onto the same counterfactual CCP – i.e., even though the model restrictions may not suffice to point identify the model parameters π, they may suffice to identify counterfactual behavior $\widetilde{p}$. This extends KSS, who have previously shown that counterfactual point-identification is achieved in the absence of additional model restrictions if and only if $rank(\mathcal{C}_J) = 0$.

3.1 Dimension Reduction

To aid practitioners, who may be interested in reducing the dimension of the identified set to obtain more informative bounds (possibly desired for policymakers), as well as for computational gains, we explore some practical alternatives. The dimension of the counterfactual identified set depends on the counterfactual transformation, the model restrictions, and the data. Dimension reduction can therefore be obtained either via the type of transformation h (e.g., based on a "localized counterfactual," discussed below) or imposing further payoff restrictions (as applied researchers rarely leave payoffs entirely unconstrained), or via an interaction of the two. In this subsection we explore "local" counterfactuals and parametric payoffs.

3.1.1 "Local Counterfactuals"

Applied researches often consider counterfacuals in which only parts of the payoff function are changed, while the rest remain unaltered – we call these "local counterfactuals." This type of counterfactual determines the structure of the $\mathcal{H}$ matrix, which in turn can affect the dimension of $\widetilde{\mathbf{P}}^I$ substantially (see equation (15)). Applied examples of local counterfactuals include entry cost subsidies (as in our numerical and Monte Carlo examples) and exports subsidies (as investigated in our empirical exercise), among several others.

Example. Consider a binary choice model $\mathcal{A} = \{0, 1\}$, and a counterfactual that only changes the payoff of one action, say $a = 1$. Take the reference action to be $J = 0$, and arrange all vectors and matrices as described in the previous sections. Then, we have

$$\mathcal{H} = \begin{bmatrix} \mathcal{H}_{11} & 0 \\ 0 & \mathcal{H}_{00} \end{bmatrix},$$

where $\mathcal{H}_{11} \neq I$ and $\mathcal{H}_{00} = I$. Consequently, $\widetilde{\pi}_1 = \mathcal{H}_{11}\pi_1 \neq \pi_1$ and $\widetilde{\pi}_0 = \mathcal{H}_{00}\pi_0 = \pi_0$ (assuming $g = 0$). In the absence of the equality restrictions (8), the dimension of $\widetilde{\mathbf{P}}^I$ equals the rank of $\mathcal{C}_0$. From (15), we obtain

$$\mathcal{C}_0 = \mathcal{H}_{11}M_1 - M_1 = [\mathcal{H}_{11} - I]M_1,$$

where $M_1 = (I - \beta F_1)(I - \beta F_0)^{-1}$. Because M_1 is invertible, it follows that

$$rank(\mathcal{C}_0) = rank(\mathcal{H}_{11} - I).$$

This rank drops to the extent that the eigenvalues of $\mathcal{H}_{11}$ equal one. In fact, if $\mathcal{H}_{11}$ is diagonalizable,

$$rank(\mathcal{C}_0) = \#\left\{\text{eigenvalues of } \mathcal{H}_{11} \text{ different from } 1\right\},$$

where $\#\{\}$ denotes the cardinality of a set. This straightforward calculation allows us to ascertain the dimension of the identified set in practice.

Our next proposition extends the example above to a more general setting.

Proposition 2. *Suppose the counterfactual transformation only changes a subset of the payoff vector π. Index the action-state pair of π by l, and partition the set of indices into the sets $\mathbb{L}$ and $\mathbb{L}'$. Assume counterfactual changes occur over the set $\mathbb{L}$, and payoffs stay the same over $\mathbb{L}'$, i.e.,*

$$\widetilde{\pi}(\mathbb{L}) = \sum_{l \in \mathbb{L}} \mathcal{H}_l(\mathbb{L}) \pi(l),$$
$$\widetilde{\pi}(\mathbb{L}') = \pi(\mathbb{L}'),$$

where $\widetilde{\pi}(\mathbb{L})$ selects all l^{th} entries of $\widetilde{\pi}$ with $l \in \mathbb{L}$ and $\widetilde{\pi}(\mathbb{L}')$ is defined similarly; $\mathcal{H}_l$ denotes the l^{th} column of $\mathcal{H}$, while $\mathcal{H}_l(\mathbb{L})$ stands for the entries of the l^{th} column with entries located in $\mathbb{L}$. Assume without loss of generality that J is counterfactual-invariant (i.e., $\widetilde{\pi}_J = \pi_J$) so that all pairs (J, x), $x \in \mathcal{X}$, belong to $\mathbb{L}'$. Then, in the absence of the equality restrictions (8), the dimension of $\widetilde{\mathbf{P}}^I(p, F)$, for $(p, F) \in \mathbf{P} \times \mathbf{F}$, is given by the rank of matrix $\mathcal{C}_J$, defined in (15), which satisfies

$$rank(\mathcal{C}_J) \leq \#\left\{\textit{eigenvalues of } \mathcal{H}(\mathbb{L}) \textit{ different from 1}\right\} \leq L,$$

where $L = \#\{\mathbb{L}\}$. The first inequality becomes an equality when $\mathcal{H}(\mathbb{L})$ is diagonalizable.

Local counterfactuals can reduce the dimension of the identified set $\widetilde{\mathbf{P}}^I$ from X to L or less, in the absence of any model restriction (8) and regardless of the observed data (p, F). If L is small, a considerable reduction occurs. Furthermore, because $\mathcal{H}$ is determined by the econometrician, it is not difficult to find the appropriate eigenvalues and, therefore, find a sharp upper bound on the dimension of $\widetilde{\mathbf{P}}^I$ in practice.

3.1.2 Parametric Payoffs

The vast majority of empirical applications rely on parametric payoffs. Here, we consider how parameterization can potentially shrink the counterfactual identified set. We assume a flexible payoff parameterization of the form

$$\pi_a = z_a \gamma + \delta_a, \tag{16}$$

for all $a \in \mathcal{A}$, where the vector of parameters γ has length η_γ, the matrix z_a has size $X \times \eta_\gamma$, and the vector δ_a has length X. Clearly, we are interested in the case where $\eta_\gamma \leq X$. Note that the specification (16) satisfies the linear restriction (8), as shown in Appendix E.

Proposition 3. *Assume the parametric payoff (16), with $\eta_\gamma \leq X$. Then, for any given counterfactual transformation h, the dimension of $\widetilde{\mathbf{P}}^I(p, F)$, for $(p, F) \in \mathbf{P} \times \mathbf{F}$, is given by $rank(\mathcal{C}_J \mathcal{Q}_J) \leq \eta_\gamma$, where $\mathcal{Q}_J = z_J$.*

Intuitively, parametric restrictions reduce the number of "free parameters" in π, which leads to dimension reduction in the counterfactual identified set $\widetilde{\mathbf{P}}^I$. This dimension reduction can be significant if the number of parameters is small. Furthermore, even when the number of parameters is not too small, it is still possible to obtain substantial dimension reduction depending on the interaction of the counterfactual transformation and the parametric assumption. That is the case when we combine a local counterfactual with the parametric restriction, as our next result shows.

Proposition 4. *Assume the parametric payoff (16), with $\eta_\gamma \leq X$, and consider a local counterfactual. Specifically, define the set of indices $\mathbb{L} \subseteq \{1, 2, ..., \eta_\gamma\}$, with cardinality $L = \#\{\mathbb{L}\}$, and let $\gamma(\mathbb{L})$ denote the subvector of γ consisting of the entries with indices in $\mathbb{L}$. Let $\mathbb{L}'$ be the set of remaining indices, and define $\gamma(\mathbb{L}')$ accordingly. The counterfactual transformation only changes $\gamma(\mathbb{L})$ and takes the following form:*

$$\begin{aligned} \widetilde{\gamma}(\mathbb{L}) &= \mathcal{D}\, \gamma(\mathbb{L}) + g(\mathbb{L}), \\ \widetilde{\gamma}(\mathbb{L}') &= \gamma(\mathbb{L}'), \end{aligned} \tag{17}$$

where $\mathcal{D}$ is an $L \times L$ matrix, $g(\mathbb{L})$ is an L vector, and both $\mathcal{D}$ and $g(\mathbb{L})$ are pre-specified by the econometrician. Then, the dimension of $\widetilde{\mathbf{P}}^I(p, F)$, for $(p, F) \in \mathbf{P} \times \mathbf{F}$, is given by

$$rank(\mathbf{M}\, Z(\mathbb{L})\, (\mathcal{D} - I)) \leq \#\, \{\textit{eigenvalues of } \mathcal{D} \textit{ different from 1}\} \leq L \leq \eta_\gamma,$$

where $Z(\mathbb{L}) = [z_{-J}(\mathbb{L})', z_J(\mathbb{L})']'$, and $z_{-J}(\mathbb{L})$ stacks $z_a(\mathbb{L})$ for all $a \neq J$, and $z_a(\mathbb{L})$ is the $X \times L$ matrix that selects all l^{th} entries of z_a with $l \in \mathbb{L}$. The first inequality becomes an equality when $\mathcal{D}$ is diagonalizable.

The upper bound on the dimension of the identified set presented in Proposition 4 is smaller than the upper bound in Proposition 3, implying that the combination of a parametric assumption with a localized counterfactual can lead to substantial dimension reduction. Concretely, when the number of *modified* parameters L is small, the dimension of $\widetilde{\mathbf{P}}^I$ is significantly reduced. (For instance, when only one parameter is changed, $L = 1$ and $\widetilde{\mathbf{P}}^I$ is one-dimensional.) Furthermore, as before, checking the dimension of the identified set in practice remains straightforward – a matter of counting the number of eigenvalues of $\mathcal{D}$ that are different from one.

4 Identification of Counterfactual Outcomes of Interest

As the state space $\mathcal{X}$ can be large in practice (making both $\widetilde{p}$ and $\widetilde{V}$ high-dimensional vectors), researchers are often interested in low-dimensional objects such as the average effects of policy interventions. Thus, in this section, our focus shifts towards characterizing, and computing, the identified set of these specific objects. Denote the low-dimensional counterfactual outcome of interest by $\theta \in \Theta \subset \mathbb{R}^n$, where Θ is a compact set (the parameter space for θ), and n is much smaller than the size of the state space X (i.e., $n \ll X$). We assume that

$$\theta = \theta(\widetilde{p}, \widetilde{s}; p, s) = \theta(p(h(s)), h(s); p(s), s). \tag{18}$$

I.e., θ depends on the counterfactual CCP, $\widetilde{p}$, and the counterfactual structure, $\widetilde{s}$, as well as on the baseline CCP, p, and the primitives, s. The second equality in (18) explicitly notes that θ depends ultimately on the baseline primitives s and the counterfactual transformation h, and it reveals the channels through which these factors can directly or indirectly affect the outcome of interest.

In our upcoming analysis, it will prove useful to supress some of the primitives in our notation, such as $\mathcal{A}$ and β in s, but retain $\widetilde{p}$ and p, even though they are themselves functions of s, and rewrite (18) as follows:

$$\theta = \phi(\widetilde{p}, p, F, \pi). \tag{19}$$

So, the counterfactual transformation h and all primitives other than F and π remain implicit and in the backgound.

Example. Projections of Counterfactual CCP: Suppose the outcome of interest consists of counterfactual choice probabilities associated with a subset of actions $\widetilde{\mathcal{A}}^* \subseteq \widetilde{\mathcal{A}}$ and states $\widetilde{\mathcal{X}}^* \subseteq \widetilde{\mathcal{X}}$. In this case, an element of θ is given by $\theta(a^*, x^*) = \widetilde{p}_{a^*}(x^*)$, for $a^* \in \widetilde{\mathcal{A}}^*$, $x^* \in \widetilde{\mathcal{X}}^*$.

Example. Average Treatment Effects: Take an outcome variable of interest that depends on actions and states, $Y_a(x, \varepsilon)$ (e.g., consumer surplus, or a firm's entry probability), with a corresponding counterfactual given by $\widetilde{Y}_a(x, \varepsilon)$. (Here, we omit the dependence of Y on all the primitives, s, to simplify notation.) The average treatment effect of the policy intervention on Y is then $\theta = \mathbb{E}[\widetilde{Y}_a(x, \varepsilon)] - \mathbb{E}[Y_a(x, \varepsilon)]$, where $\mathbb{E}[\widetilde{Y}_a(x, \varepsilon)]$ integrates over the distribution of actions and states in the counterfactual scenario, while $\mathbb{E}[Y_a(x, \varepsilon)]$ integrates over the factual distribution. One may consider the long-run distribution, or may condition on an initial state and estimate short-run effects.

Our next proposition follows.

Proposition 5. *The sharp identified set for θ is*

$$\mathbf{\Theta}^I(p,F) = \left\{ \begin{array}{c} \theta \in \Theta : \exists\, (\widetilde{p},\pi) \in \mathbf{P} \times \mathbb{R}^{(A+1)X} \text{ such that} \\ \theta = \phi\left(\widetilde{p}, p, F, \pi\right), \; \mathbf{M}(F)\pi = b_{-J}(p,F), \\ R^{eq}\pi = r^{eq}, \;\; R^{iq}\pi \leq r^{iq}, \\ (\widetilde{\mathbf{M}}(F)\mathcal{H})\,\pi = \widetilde{b}_{-J}\left(\widetilde{p},F\right) - \widetilde{\mathbf{M}}(F)g \end{array} \right\}, \tag{20}$$

for $(p,F) \in \mathbf{P} \times \mathbf{F}$. When ϕ is a continuous function of $(\widetilde{p},\pi)$, $\mathbf{\Theta}^I(p,F)$ is a connected set. In addition, when $\Pi^I(p,F)$ is bounded, $\mathbf{\Theta}^I(p,F)$ is compact. Finally, if θ is a scalar, then $\mathbf{\Theta}^I(p,F)$ is an interval.

Proposition 5 states that a vector θ belongs to $\mathbf{\Theta}^I$ if and only if there exists a payoff π that is compatible with the data (i.e., $\mathbf{M}(F)\,\pi = b_{-J}(p,F)$), satisfies the model restrictions (i.e., $R^{eq}\pi = r^{eq}$ and $R^{iq}\pi \leq r^{iq}$), can generate $\widetilde{p}$ in the counterfactual scenario (i.e., $(\widetilde{\mathbf{M}}(F)\mathcal{H})\,\pi = \widetilde{b}_{-J}\left(\widetilde{p},F\right) - \widetilde{\mathbf{M}}(F)g$), and the corresponding pair $(\widetilde{p},\pi)$ can generate θ (i.e., $\theta = \phi\left(\widetilde{p},p,F,\pi\right)$).

When ϕ is continuous, $\mathbf{\Theta}^I$ is connected because it is the image set of a (composite) continuous function defined on the convex polyhedron Π^I. If the model restrictions make Π^I bounded, $\mathbf{\Theta}^I$ becomes a compact and connected set, which is convenient as it suffices to trace the boundary of $\mathbf{\Theta}^I$ to characterize this set in practice. In addition, when θ is a scalar, $\mathbf{\Theta}^I$ reduces to a compact interval, which is even simpler to characterize: in that case we just need to compute the lower and upper endpoints of the interval $\mathbf{\Theta}^I$.

The upper endpoint of this interval can be calculated by solving the following constrained maximization problem:

$$\theta^U \equiv \max_{(\widetilde{p},\pi)\in\widetilde{\mathbf{P}}\times\mathbb{R}^{(A+1)X}} \phi\left(\widetilde{p},p,F,\pi\right) \tag{21}$$

subject to

$$\begin{array}{rcl} \mathbf{M}(F)\,\pi & = & b_{-J}(p,F), \\ R^{eq}\,\pi & = & r^{eq}, \\ R^{iq}\,\pi & \leq & r^{iq}, \\ (\widetilde{\mathbf{M}}(F)\mathcal{H})\,\pi & = & \widetilde{b}_{-J}\left(\widetilde{p},F\right) - \widetilde{\mathbf{M}}(F)g. \end{array} \tag{22}$$

The lower bound of the identified set θ^L is defined similarly (but replacing max by min).[16]

Computation of $\mathbf{\Theta}^I$. The problem (21)–(22) is a nonlinear maximization problem with linear constraints on π and smooth nonlinear constraints on $\widetilde{p}$, taking the data (p,F) as given. When ϕ is differentiable, the optimization is well-behaved and can be solved using standard software (e.g., Knitro), even when the state space is large.

[16] In Appendix G, we provide a discussion on how to extend (21)–(22) to incorporate an unknown discount factor β and distribution G, as well as the associated practical challenges with such an extension.

In our experience, standard algorithms are highly efficient in solving (21)–(22) in empirically-relevant high-dimensional problems when the researcher provides the gradient of ϕ (see our Monte Carlo simulations in Appendix D). In some cases, however, the gradient of ϕ may be nontrivial to compute; for instance, this is the case when the target parameter θ involves average effects based on both factual and counterfactual ergodic distributions of the states. For such cases, we show in Appendix H how to calculate the gradient of ϕ analytically to help the numerical search.

In other cases, numerical gradients are costly to evaluate, and then standard solvers can be slow to converge. We thus develop a new a stochastic algorithm that exploits the structure of the problem (21)–(22) and combines the strengths of alternative stochastic search procedures. We discuss and describe our proposed algorithm in Appendix B.[17]

5 Example: Firm Entry/Exit Model

To fix ideas, we illustrate the identified sets in the context of a simple firm entry/exit problem. Suppose firm i faces the choice set $\mathcal{A} = \{\text{out}, \text{in}\} = \{0, 1\}$. Decompose the state space into $x_{it} = (k_{it}, w_{it})$, where $k_{it} \in \mathcal{K} = \{0, 1\}$ is the lagged decision a_{it-1}, and $w_{it} \in \mathcal{W} = \{1, ..., W\}$ is an exogenous profit shifter (e.g. market size). Assume for convenience that w_{it} can take two values, low and high: $\mathcal{W} = \{w^l, w^h\}$, with $w^l < w^h$. The size of the state space is therefore $X = KW = 4$, where $K = \#\{\mathcal{K}\}$ and $W = \#\{\mathcal{W}\}$. Transition probabilities are decomposed as $F(k_{it+1}, w_{it+1}|k_{it}, w_{it}, a) = F(k_{it+1}|k_{it}, a)F(w_{it+1}|w_{it})$.

Let $\pi_a(k)$ denote the $W \times 1$ payoff vector the firm obtains when it chooses action a given k and w, so that $\pi_a = [\pi_a'(0), \pi_a'(1)]'$. We impose the following structure on π:

$$\pi_0 = \begin{bmatrix} oo \\ s \end{bmatrix}, \pi_1 = \begin{bmatrix} vp - fc - ec \\ vp - fc \end{bmatrix}. \tag{23}$$

The payoff the firm obtains when it was out of the market in the previous period and stays out in the current period is the vector $\pi_0(0) = oo$ (the value of the outside option); and the payoff when the firm was active and decides to exit is given by the vector of scrap values, $\pi_0(1) = s$. Note that both the outside option and the scrap values can vary with the exogenous state w. The vectors vp, fc, and ec are the variable profits, the fixed costs, and the entry costs, respectively (all of which can vary with w as

[17] Intuitively, one can search for θ^U by finding admissible values and updated directions for π, which is not computationally difficult as it only depends on linear constraints. But finding the corresponding $\widetilde{p}$ by solving the nonlinear equation (13) repeatedly (and calculating the gradient of ϕ numerically) can be demanding. Alternatively, searching stochastically over $\widetilde{p}$ and then finding a compatible π satisfying linear constraints is simpler. However, $\widetilde{\mathbf{P}}^I$ may be a "thin" set in $\widetilde{\mathbf{P}}$ with an unknown shape (since its dimension can be much smaller than the dimension of the simplex; see Proposition 1 and the results in Section 3.1). Consequently, it is difficult to find points within that set randomly, and it is easy for perturbation methods to "exit" the set, increasing the cost of finding the maximum θ. We therefore develop a new algorithm in which we move in the "$\widetilde{p}$-world" (to avoid solving (13) repeatedly), but we keep a close eye on the "π-world" (to keep track of the model restrictions and search in relevant directions). Searching in relevant directions without solving (13) and computing the numerical gradient of ϕ in every step improves substantially how fast θ moves on each iteration to the maximum. See Appendix B for details.

well). The vector $\pi_1(0) = vp - fc - ec$ measures the profits the firm gets when it enters the market, and $\pi_1(1) = vp - fc$ are the profits when it stays.

In this example, both π_0 and π_1 are 4×1 vectors (and so π has $2X = 8$ elements). To point-identify π we need $X = 4$ restrictions. Typically, researchers identify an entry model by setting $oo = 0$ (for 2 restrictions); and further setting either $s = 0$ or assuming $vp - fc$ is known (e.g., by assuming variable profits vp can be recovered "offline," using price and quantity data, and setting $fc = 0$). When $oo = s = 0$, then $\pi_0 = 0$, and point identification of π follows directly from (2); it is essentially a restriction on a reference action. When instead $\pi_0(0) = oo = 0$ and $vp - fc$ is known, we identify the remaining elements of π by combining (2) and (8).

Assuming the outside option equals the scrap value or the fixed costs (and all are equal to zero) may be difficult to justify in practice, as cost or scrap value data are extremely rare (Kalouptsidi, 2014). When the researcher is not willing to impose such restrictions, π is not point-identified. Yet, the payoff function can be set-identified under weaker conditions. Consider, for instance, the following set of assumptions:

1. $oo = 0, \quad fc \geq 0, \quad ec \geq 0$, and vp is known.
2. $\pi_1(1, w^h) \geq \pi_1(1, w^l)$, and $vp - fc \leq ec \leq \frac{\mathbb{E}[vp-fc]}{1-\beta}$, where the expectation is taken over the ergodic distribution of the state variables.
3. s does not depend on w.

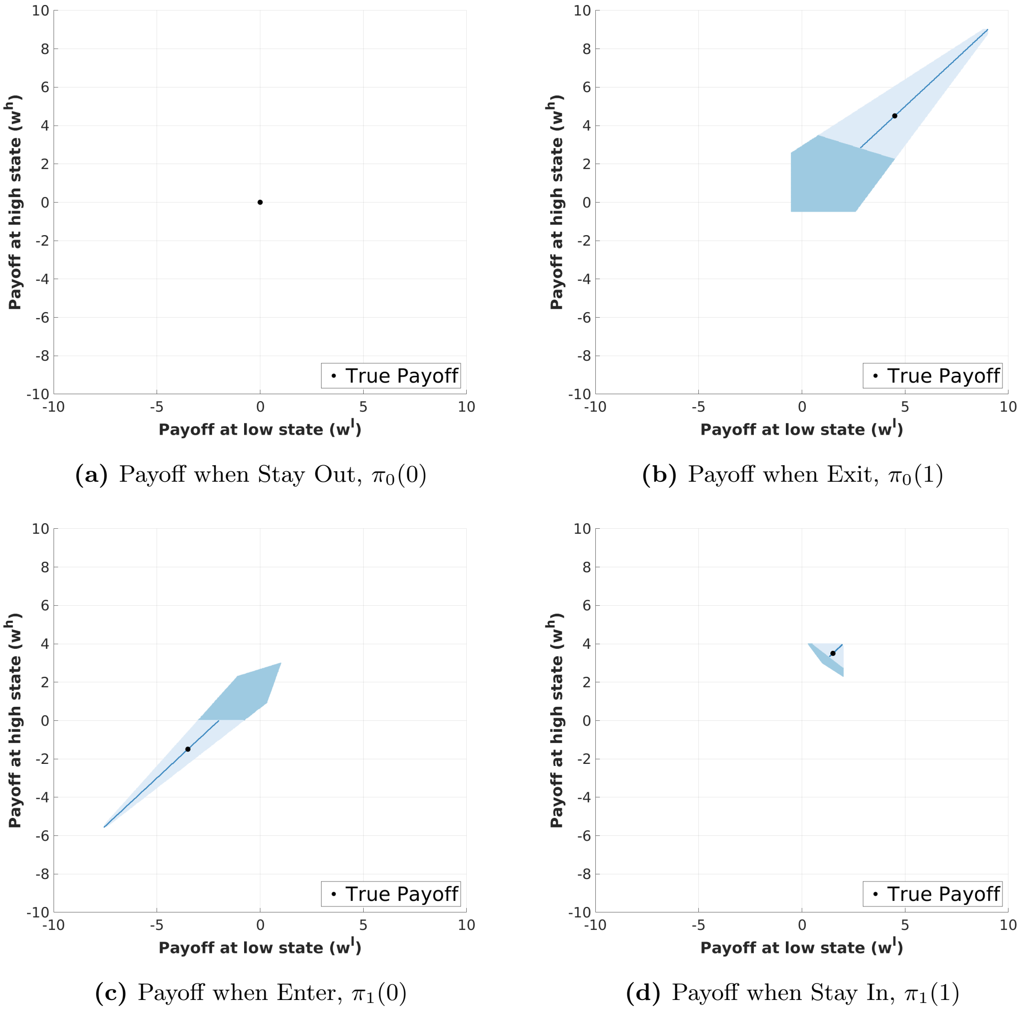
Restriction 1 assumes that the outside option is zero (as usual); fixed costs and entry costs are both positive; and variable profits are known (estimated "offline"). This set of restrictions imposes $d = W = 2$ equality and $m = 4$ inequality constraints. From (10), it is clear that the identified set Π^I is a two-dimensional set ($X - d = 2$) in the eight-dimensional space.

Restriction 2 imposes $m = 5$ inequality constraints: profits are increasing in w when the firm is in the market (a monotonicity assumption); entry costs are greater than variable profits minus fixed costs (implying that entry is always costly in the first period of entry); and ec is smaller than the expected present value of future profits when the firm stays forever in the market (meaning that, on average, it eventually pays off to enter).

Restriction 3 assumes an exclusion restriction: scrap values are state-invariant. This corresponds to $d = W - 1 = 1$ equality restriction. Note that, by combining Restrictions 1 and 3, we obtain $d = 3$ linear equalities, which makes the identified set Π^I one dimensional. In Appendix F, we provide explicit characterizations for this example.

Figure 1 presents the identified set for payoffs, Π^I, for a particular parameter configuration.[18] The

[18] We assume scrap values, entry and fixed costs do not depend on w and take the following values: $s = 4.5$, $ec = 5$, and $fc = 0.5$. We also impose $vp(w^l) = 2$ and $vp(w^h) = 4$, so that $\pi_0 = (0, 0, 4.5, 4.5)'$ and $\pi_1 = (-3.5, -1.5, 1.5, 3.5)'$. The discount factor is $\beta = 0.9$, the transition process for w is $Pr(w_{t+1} = w^l | w_t = w^l) = Pr(w_{t+1} = w^h | w_t = w^h) = 0.75$, and the idiosyncratic shocks ε_{it} follow a type 1 extreme value distribution (the scale parameter is set at $\sigma = 1$). Under these assumptions, $\frac{\mathbb{E}[vp-fc]}{1-\beta} = 25$.



(a) Payoff when Stay Out, $\pi_0(0)$

(b) Payoff when Exit, $\pi_0(1)$

(c) Payoff when Enter, $\pi_1(0)$

(d) Payoff when Stay In, $\pi_1(1)$

Figure 1: Firm Entry/Exit Model: Payoff Identified Set Π^I under Restrictions 1, 2, and 3. The larger polyhedron (including the dark blue areas) correspond to Π^I under Restriction 1. The light blue areas correspond to Π^I under Restrictions 1 and 2. The identified set Π^I under Restrictions 1–3 is represented by the blue lines within the light blue polyhedron. The true π is represented by the black dots.

larger polyhedron corresponds to Π^I under Restriction 1. The identified set is informative despite the fact that the assumptions imposed are not overly restrictive. For a brief intuition of how the linear (in)equalities interact to produce Figure 1, consider the set corresponding to scrap values (panel (b)). In this model, equation (7) alone implies that the difference between scrap values and entry costs is point-identified (see Appendix F). As a consequence, the inequality $ec \geq 0$ implies a lower bound on scrap values (for each state w), shifting the origin. Similarly, equation (7) implies that the sum of scrap values and the present value of fixed costs is point-identified. The inequality $fc \geq 0$ entails an upper bound on scrap values, eliminating from the identified set all values for s above the intersecting lines shown in the figure. The increasing lines reflect the fact that equation (7) relates s and the *present value* of fc, so that $fc \geq 0$ leads to restrictions on scrap values *across* states.

Restrictions 1 and 2 together lead to substantial identifying power: Π^I now corresponds to the light blue polyhedron, which is substantially smaller. Assuming that entry is costly in the first period of entry, $vp - fc \leq ec$, is the main restriction responsible for the reduction in the identified set. This assumption results in another lower bound on s (see panel (b)), but differently from $ec \geq 0$, it involves restrictions on fc and so imposes restrictions on s across states; the other assumptions in Restriction 2 are not as informative in this example; see Appendix F. Interestingly, the payoff function with scrap values that are equal to zero does not belong to Π^I under these two sets of restrictions. As mentioned previously, setting scrap values to zero is a common way to point-identify π, but, given that $s = 0$ is at odds with Restrictions 1 and 2, such assumption would be rejected by the data.

Finally, Restriction 3 (exclusion restriction on scrap values) also has substantial identifying power as it reduces the dimension of the identified set to one. In the figures, the identified set under Restrictions 1–3 is represented by the blue lines within the light blue polyhedron.

Counterfactuals. The counterfactual experiment we consider is a subsidy that decreases entry costs by 20% – a "local counterfactual." Formally, $\widetilde{\pi} = \mathcal{H}\pi + g$, with $g = 0$, and $\mathcal{H}$ block-diagonal with the diagonal blocks $\mathcal{H}_{00}$ and $\mathcal{H}_{11}$ given by

$$\mathcal{H}_{00} = I, \text{ and } \mathcal{H}_{11} = \begin{bmatrix} \tau I & (1-\tau)I \\ 0 & I \end{bmatrix},$$

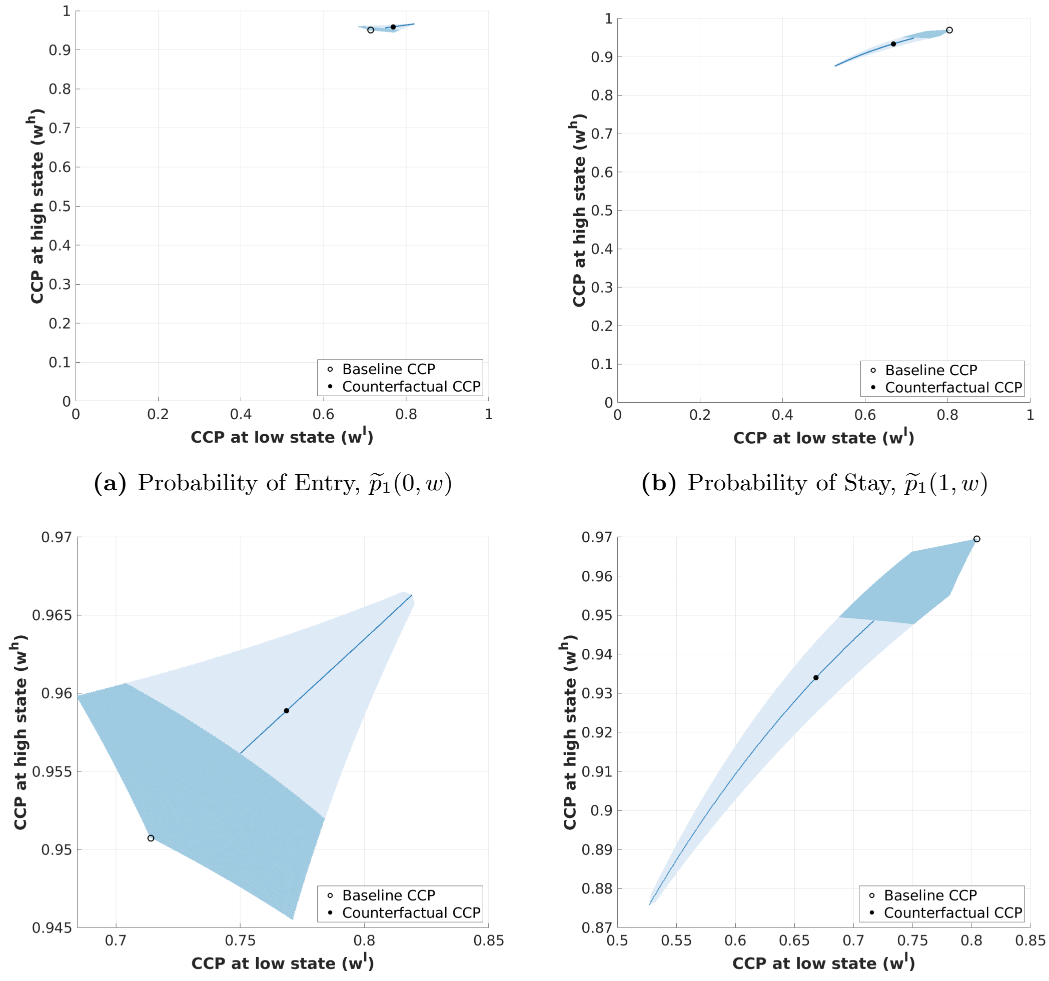
where $\tau = 0.8$. This implies

$$\widetilde{\pi}_0 = \mathcal{H}_{00}\pi_0 = \pi_0, \text{ and } \widetilde{\pi}_1 = \mathcal{H}_{11}\pi_1 = \begin{bmatrix} vp - fc - \tau \times ec \\ vp - fc \end{bmatrix}.$$

It is clear from the outset that the identified set $\widetilde{\mathbf{P}}^I$ is two-dimensional, given that $\mathcal{H}_{11}$ is diagonalizable with two eigenvalues that are different from one (see Proposition 2).

Figure 2 presents the results. First, note that the baseline and counterfactual CCPs, p and $\widetilde{p}$, are represented by the black empty circle and the black full dot, respectively. In the baseline scenario, there is a higher probability of entering (and staying) in the high state than in the low state because higher values of w lead to greater profits and because w follows a persistent Markov process. In the counterfactual, the subsidy increases the probability of entry compared to the baseline in both low and high states w (as it should). Moreover, the subsidy decreases the probability of staying in the market, as it becomes cheaper to re-enter in the future.

We now characterize the identified set $\widetilde{\mathbf{P}}^I$ under Restrictions 1–3. Similar to our representation of Π^I in Figure 1, the larger sets (including the dark blue areas) correspond to $\widetilde{\mathbf{P}}^I$ under Restriction 1. The identified set is highly informative: it is a two-dimensional set in a four-dimensional space, as noted earlier, excluding most points in $\widetilde{\mathbf{P}}$ from being possible counterfactual CCPs. Yet, because the baseline CCP p



(a) Probability of Entry, $\widetilde{p}_1(0, w)$

(b) Probability of Stay, $\widetilde{p}_1(1, w)$

(c) Zoom: Probability of Entry, $\widetilde{p}_1(0, w)$

(d) Zoom: Probability of Stay, $\widetilde{p}_1(1, w)$

Figure 2: Identified Set for Counterfactual CCPs, $\widetilde{\mathbf{P}}^I$, under Restrictions 1, 2, and 3. The larger sets (including the dark blue areas) correspond to $\widetilde{\mathbf{P}}^I$ under Restriction 1. The light blue areas correspond to $\widetilde{\mathbf{P}}^I$ under Restrictions 1 and 2. The identified set $\widetilde{\mathbf{P}}^I$ under Restrictions 1–3 is represented by the blue lines within the light blue areas. The baseline and counterfactual CCPs, p and $\widetilde{p}$, are represented by the black empty circle and the black full dot, respectively. The bottom panels present the "zoomed-in" versions of the top panels.

is at the boundary of $\widetilde{\mathbf{P}}^I$, one cannot rule out the possibility that the entry subsidy has no impact on the firm's behavior. Adding Restriction 2 reduces the size of $\widetilde{\mathbf{P}}^I$ substantially (corresponding to the light blue areas in the figure). This is a direct consequence of the smaller set Π^I after imposing Restrictions 1 and 2 (see Figure 1). The baseline CCP does not belong to $\widetilde{\mathbf{P}}^I$ once we add Restriction 2; in fact, the location of p and $\widetilde{\mathbf{P}}^I$ allows us to conclude that the probability of entry increases in the counterfactual and that the probability of staying decreases. In other words, the sign of the treatment effect is identified. The exclusion restriction on scrap values (Restriction 3) has substantial identification power, making $\widetilde{\mathbf{P}}^I$ one-dimensional (because Π^I becomes one-dimensional as well) – see the blue lines in the figure. Note that all identified sets are connected, as expected (Proposition 1), but not necessarily convex.

We now turn to some low-dimensional outcomes θ, in particular, the long-run average impact of the entry subsidy on (i) the probability of staying in the market (labelled θ_P), (ii) consumer surplus (θ_S), and (iii) the value of the firm (θ_V). Table 1 presents the identified sets for each of these outcomes under Restrictions 1–3.[19]

Table 1: Sharp Identified Sets for the Long-run Impact of the Entry Subsidy on Outcomes of Interest, $\mathbf{\Theta}^I$

Outcome of Interest	Target parameter	Sharp Identified Sets		
	True	Restriction 1	Restrictions 1–2	Restrictions 1–3
Change in Prob. of Being Active	-0.0638	[-0.1235, 0.0000]	[-0.1235, -0.0341]	[-0.1235, -0.0421]
	True	Restriction 1	Restrictions 1–2	Restrictions 1–3
Change in Consumer Surplus	-0.0875	[-0.1735, 0.0000]	[-0.1735, -0.0474]	[-0.1735, -0.0573]
	True	Restriction 1	Restrictions 1–2	Restrictions 1–3
Change in the Value of the Firm	0.9513	[0.0000, 1.8229]	[0.4489, 1.8229]	[0.6388, 1.8229]

Notes: This table shows the true parameter, as well as the sharp identified sets for the long-run average effect of the 20% entry subsidy on three outcomes of interest in the firm entry/exit problem: the probability of staying active, the consumer surplus, and the value of the firm. The averages are taken with respect to the state variables, using the steady-state distribution. See Appendix F for details.

Perhaps surprisingly, the entry subsidy decreases the long-run average probability of the firm staying in the market, by approximately 6.4 percentage points. That is because, while the subsidy induces more entry, it also induces more exit. In the current case, increasing both firm's entry and exit rates results in less time spent in the market in the long run. This in turn reduces the long-run average consumer surplus, and raises the average long-run value of the firm.

As expected, the identified sets are all compact intervals (Proposition 5), and they all contain the true θ. Under Restriction 1, the upper bound of the identified set for θ_P is zero, leading to the conclusion that the long-run average probability of being active does not increase in the counterfactual. The lower bound implies that the probability of staying active can be reduced by at most 12 percentage points. Similarly, the researcher can conclude that the long-run average consumer surplus does not go up (and decreases by at most \$0.17), while the long-run average value of the firm does not go down (and increases at most by \$1.8) in response to the subsidy. These are informative identified sets despite the fact that Restriction 1 is mild.

Adding Restriction 2 makes all identified sets more informative. The upper bound on θ_P is now lower, implying that the average probability of being active is now reduced by a number between 3.4 and 12 percentage points, which clearly identifies the sign of the impact. The endpoints of the intervals for

[19] Assuming a (residual) linear inverse demand $P_{it} = w_{it} - \eta Q_{it}$, where P_{it} is the price and Q_{it} is the quantity demanded, and assuming a constant marginal cost mc, the variable profit is given by $vp = (w_{it} - mc)^2/4\eta$. The consumer surplus is $S = 0$ when the firm is inactive ($a = 0$), and $S = (w_{it} - mc)^2/8\eta$ when it is active ($a = 1$). The value of the firm in the baseline is given by the vector $V = (I - \beta F_J)^{-1}\left(\pi_J + \psi_J(p)\right)$, where we take $J = 0$ (see footnote 10), and a similar expression holds for the counterfactual value: $\widetilde{V} = (I - \widetilde{\beta}\widetilde{F}_J)^{-1}(\widetilde{\pi}_J + \widetilde{\psi}_J(\widetilde{p}))$. See Appendix F for explicit formulas for $\theta = (\theta_P, \theta_S, \theta_V)$.

θ_S and θ_V change similarly. Adding Restriction 3 does not narrow the intervals much further, despite the fact that this restriction has substantial identifying power related to the model parameters π and counterfactual behavior $\widetilde{p}$. That is because, while Restriction 3 reduces the dimension of Π^I and $\widetilde{\mathbf{P}}^I$, it does not affect substantially the extreme points of these sets that in turn generate the endpoints of $\mathbf{\Theta}^I$. In Appendix F, we present the three-dimensional identified set $\mathbf{\Theta}^I$.

In Appendix D, we present results from a Monte Carlo study based on this example of firm entry and exit. Our findings there are analogous: the sets are informative even under the mildest restrictions and always contain the true parameter values. Moreover, calculating the bounds for θ is computationally fast, even in cases where the state space is large.

6 Estimation and Inference

We now present the inference procedure for the main outcomes of interest θ. In particular, we want to construct confidence sets (CS's) for the true value of θ (rather than for the identified set $\mathbf{\Theta}^I$). Our approach is similar in spirit to the Hotz and Miller (1993) two-step estimator: we estimate choice probabilities p and transitions of state variables F in the first step, and then we perform inference on θ in the second step.

We assume the econometrician has access to a panel data on agents' actions and states: $\{a_{it}, x_{it} : i = 1, ..., N;\ t = 1, ..., T\}$. We consider asymptotics for the large N and fixed T case, as is typical in microeconometric applications, and assume i.i.d. sampling in the cross-section dimension.[20] Given that actions and states are finite, we consider frequency estimators for both p and F. Specifically, for all $a \in \mathcal{A}$, and all $x, x' \in \mathcal{X}$,

$$\widehat{p}_{aN}(x) = \frac{\sum_{it} 1\{x_{it} = x, a_{it} = a\}}{\sum_{it} 1\{x_{it} = x\}}, \tag{24}$$

$$\widehat{F}_{aN}(x', x) = \frac{\sum_{it} 1\{x_{it+1} = x', x_{it} = x, a_{it} = a\}}{\sum_{it} 1\{x_{it} = x, a_{it} = a\}}, \tag{25}$$

and the vectors of sample frequencies are denoted by $\widehat{p}_N$ and $\widehat{F}_N$.[21] We collect the terms $\widehat{p}_N$ and $\widehat{F}_N$ into the L–vector $\widehat{\mathfrak{p}}_N = [\widehat{\mathfrak{p}}_{1N}, ..., \widehat{\mathfrak{p}}_{LN}]'$. Similarly, we collect p and F into $\mathfrak{p} = [\mathfrak{p}_1, ..., \mathfrak{p}_L]' := E[\mathbf{e}]$, where $\mathbf{e}$ is a vector of observed indicators. Recall that each matrix $M_a, a \in \mathcal{A}$, is a function of F, which is a subvector of $\mathfrak{p}$, therefore we define $M_a(\mathfrak{p}), a \in \mathcal{A}$, as the value of M_a evaluated at $\mathfrak{p}$ and also define $\mathbf{M}(\mathfrak{p})$ accordingly. We use the same notation for $b_{-J}(\mathfrak{p})$, as well as for $\widetilde{\mathbf{M}}(\mathfrak{p})$, $\widetilde{b}_{-J}(\widetilde{p}, \mathfrak{p})$, and $\phi(\widetilde{p}, \pi; \mathfrak{p})$ when appropriate.

We construct a confidence set by inverting a test. To test the null $H_0 : \theta = \theta_0$ against the alternative

[20] If the data is ergodic and an appropriate mixing condition is satisfied then our procedure remains valid when $T \to \infty$ and N is fixed.

[21] In certain cases some elements of the transition matrix F are degenerate when the corresponding states are known to evolve deterministically; see equation (F1) in the Online Appendix. We do not estimate these elements, and thus the expressions in (25) are applied only to the rest of the elements of F that need to be estimated.

$H_1 : \theta \neq \theta_0$, we reformulate the problem in the following way. For a fixed value $\theta = \theta_0$, we take the following equality constraints on π:

$$R^{eq}\pi = r^{eq},\ (\widetilde{\mathbf{M}}(\mathfrak{p})\mathcal{H})\pi = \widetilde{b}_{-J}\,(\widetilde{p},\mathfrak{p}) - \widetilde{\mathbf{M}}(\mathfrak{p})g,\ \text{and } \theta_0 = \phi(\widetilde{p},\pi;\mathfrak{p}),\ \text{for some } \widetilde{p},$$

and collect them into

$$\mathcal{R}(\theta_0,\pi,\widetilde{p};\mathfrak{p}) = 0.$$

This leads to the criterion function

$$J(\theta_0) := \min_{\substack{(\widetilde{p},\pi)\in\widetilde{\mathbf{P}}\times\mathbb{R}^{(A+1)X} \,:\, R^{iq}\pi\le r^{iq},\\ \mathcal{R}(\theta_0,\pi,\widetilde{p};\mathfrak{p})=0}} [b_{-J}(\mathfrak{p}) - \mathbf{M}(\mathfrak{p})\pi]'\,\Omega\,[b_{-J}(\mathfrak{p}) - \mathbf{M}(\mathfrak{p})\pi], \tag{26}$$

where Ω is a (user-chosen) positive definite weighting matrix. If θ_0 belongs to $\mathbf{\Theta}^I$ then all restrictions are satisfied and $J(\theta_0) = 0$, otherwise $J(\theta_0) > 0$. The identified set $\mathbf{\Theta}^I$ can therefore be represented as the set of $\theta's$ in Θ such that $J(\theta) = 0$. This implies that the null $H_0 : \theta = \theta_0$ is equivalent to $H_0' : J(\theta_0) = 0$. Reformulating the problem in this way has benefits that we discuss shortly.

The test statistic is based on the empirical counterpart of $J(\theta_0)$, which is given by

$$\widehat{J}_N(\theta_0) := \min_{\substack{(\widetilde{p},\pi)\in\widetilde{\mathbf{P}}\times\mathbb{R}^{(A+1)X} \,:\, R^{iq}\pi\le r^{iq},\\ \mathcal{R}(\theta_0,\pi,\widetilde{p};\widehat{\mathfrak{p}}_N)=0}} [b_{-J}(\widehat{\mathfrak{p}}_N) - \widehat{\mathbf{M}}_N\pi]'\,\widehat{\Omega}_N\,[b_{-J}(\widehat{\mathfrak{p}}_N) - \widehat{\mathbf{M}}_N\pi], \tag{27}$$

where $\widehat{\mathbf{M}}_N = \mathbf{M}(\widehat{\mathfrak{p}}_N)$, and $\widehat{\Omega}_N$ is a consistent estimator for Ω. For the rest of the paper we consider a general specification of Ω so that it can be a (known) continuous function of $\mathfrak{p}$. Denoting the function by $\Omega(\cdot)$, we let $\widehat{\Omega}_N = \Omega(\widehat{\mathfrak{p}}_N)$ in (27).

The rejection region of the test with significance level α is $N\widehat{J}_N(\theta_0) > \widehat{c}_{1-\alpha}(\theta_0)$, where $\widehat{c}_{1-\alpha}(\theta_0)$ is a data-dependent critical value. While a naive bootstrap for $\widehat{J}_N(\theta_0)$ fails to deliver critical values that are asymptotically uniformly valid (see, e.g. Kitamura and Stoye, 2018), subsampling works under weak conditions, as we shall show shortly. Let h_N be the subsample size, with $h_N \to \infty$ as $N \to \infty$. A subsample version of $\widehat{J}_N(\theta_0)$ is

$$\widehat{J}^*_{h_N}(\theta_0) := \min_{\substack{(\widetilde{p},\pi)\in\widetilde{\mathbf{P}}\times\mathbb{R}^{(A+1)X} \,:\, R^{iq}\pi\le r^{iq},\\ \mathcal{R}\left(\theta_0,\pi,\widetilde{p};\widehat{\mathfrak{p}}^*_{h_N}\right)=0}} [\widehat{b}^*_{-J} - \widehat{\mathbf{M}}^*_{h_N}\pi]'\widehat{\Omega}^*_{h_N}[\widehat{b}^*_{-J} - \widehat{\mathbf{M}}^*_{h_N}\pi], \tag{28}$$

where $\widehat{\mathfrak{p}}^*_{h_N}$ is a subsample estimator of $\mathfrak{p}$, $\widehat{\mathbf{M}}^*_{h_N} = \mathbf{M}(\widehat{\mathfrak{p}}^*_{h_N})$, $\widehat{\Omega}^*_{h_N} = \Omega(\widehat{\mathfrak{p}}^*_{h_N})$ and

$$\widehat{b}^*_{-J} = b_{-J}(\widehat{\mathfrak{p}}^*_{h_N}) - b_{-J}(\widehat{\mathfrak{p}}_N) + \widehat{b}_{-J}(\widehat{\mathfrak{p}}_N), \tag{29}$$

with $\widehat{b}_{-J}(\widehat{\mathfrak{p}}_N)$ being the value of $\widehat{\mathbf{M}}_N\pi$ solving the minimization problem (27). Note that with this definition of $\widehat{b}^*_{-J}$ we implement subsampling with centering.[22]

The testing procedure is simple: We use the empirical distribution of $h_N \widehat{J}^*_{h_N}(\theta_0)$ to obtain the critical value $\widehat{c}_{1-\alpha}(\theta_0)$. When the value of the test statistic is smaller than the critical value, $N\widehat{J}_N(\theta_0) \leq \widehat{c}_{1-\alpha}(\theta_0)$, we do not reject the null $H_0' : J(\theta_0) = 0$, otherwise we reject it. The $1-\alpha$ confidence set will be the collection of θ_0's for which the tests do not reject the null; in Subsection 6.1 below, we discuss the practical implementation.

The next theorem follows:

Theorem 1. *Under Condition 1 presented in Appendix A,*

$$\liminf_{N\to\infty} \inf_{(\mathfrak{p},\theta)\in\mathcal{P}} \Pr\{N\widehat{J}_N(\theta) \leq \widehat{c}_{1-\alpha}(\theta)\} = 1-\alpha,$$

where $\widehat{c}_{1-\alpha}(\theta)$ is the $1-\alpha$ quantile of $h_N \widehat{J}^*_{h_N}(\theta)$, with $0 \leq \alpha \leq \frac{1}{2}$, and, for some positive constants c_1 and c_2,

$$\mathcal{P} := \left\{ \begin{array}{c} (\mathfrak{p},\theta) : \mathfrak{p}_\ell \in (0,1), E\left[\left|\frac{\mathbf{e}_\ell}{\sqrt{\mathfrak{p}_\ell(1-\mathfrak{p}_\ell)}}\right|^{2+c_1}\right] < c_2, 1 \leq \ell \leq L, \theta \in \Theta, \\ \exists(\widetilde{p},\pi) \in \widetilde{\mathbf{P}} \times \mathbb{R}^{(A+1)X} \text{ such that } \mathbf{M}(\mathfrak{p})\pi = b_{-J}(\mathfrak{p}), \\ R^{iq}\pi \leq r^{iq},\ \mathcal{R}(\theta,\pi,\widetilde{p};\mathfrak{p}) = 0, \det(\Omega(\mathfrak{p})) \geq c_1 \end{array} \right\}.$$

The asymptotically uniformly valid $1-\alpha$ confidence set for θ is $CS = \{\theta \in \Theta : N\widehat{J}_N(\theta) \leq \widehat{c}_{1-\alpha}(\theta)\}$.

The set $\mathcal{P}$ imposes regularity conditions on the data generating process for every counterfactual value of interest $\theta \in \Theta$.[23] Our test statistic (27) is the squared minimum distance between the random vector $b_{-J}(\widehat{\mathfrak{p}}_N)$ and a random manifold. It is therefore crucial to take sampling uncertainty in both objects into account. Importantly, the manifold does not have to be convex. We avoid such standard convexity conditions as they are typically incompatible with our model restrictions, in particular the general equality restrictions $\mathcal{R}(\theta_0,\pi,\widetilde{p};\mathfrak{p}) = 0$. Theorem 1 establishes the asymptotic validity of our procedure, addressing these issues.[24]

Remark 1. We note that potentially there exist many possible choices for the criterion function $J(\cdot)$ and thus its sample counterpart $\widehat{J}_N(\cdot)$. The form of $J(\cdot)$ in the paper is motivated by our method of inference, which is based on a Chernoff-type test, with its critical values obtained by (recentered) subsampling. The Chernoff test is one of the most studied procedures involving non-regularity (see, for example, Geyer

[22] The uncentered version takes $\widehat{b}^*_{-J} = b_{-J}(\widehat{\mathfrak{p}}^*_{h_N})$, and it is asymptotically valid as well. However, in our numerical experience, it has worse finite sample behavior than the centered version.

[23] The first restriction in the definition of $\mathcal{P}$ is a standard condition imposed to guarantee the Lindeberg condition. The second, the third, and the fourth collect the model restrictions (the equalities in the fourth restriction include the constraints that arise as we fix the value of the counterfactual θ). The final restriction guarantees that the minimizations (27) and (28) are asymptotically well-behaved.

[24] Formally, the test statistic projects $b_{-J}(\widehat{\mathfrak{p}}_N)$ on the manifold $S(\widehat{\mathfrak{p}}_N,\theta_0)$, where $S(\mathfrak{p},\theta) := \{\mathbf{M}(\mathfrak{p})\pi, \pi \in \mathbb{R}^{(A+1)X} : \mathcal{R}(\theta,\pi,\widetilde{p};\mathfrak{p}) = 0$, and $R^{iq}\pi \leq r^{iq}$ hold for some $\widetilde{p} \in \widetilde{\mathbf{P}}\}$. Condition 1 does *not* require that $S(\widehat{\mathfrak{p}}_N,\theta_0)$ be convex; it requires instead that the tangent cone of $S(\widehat{\mathfrak{p}}_N,\theta_0)$ be convex (see Appendix A for details).

(1994), Shapiro (1985) and Chernoff (1954), just to name a few), and our proof strategy relies on existing results in the area. Of course it is by all means possible that one may be able to establish asymptotic uniform validity in other formulations, and that is an important topic of future investigation. We also note that our formulation accommodates efficient computational approaches as outlined in Section 6.1. We demonstrate, through Monte Carlo (Appendix D) and an empirical application (Section 7), that our algorithms are highly effective in problems of realistic and relevant scales.

Remark 2. A comment on some approaches that are alternative to ours as outlined above is in order. One such an alternative would start with treating our problem as a system of equalities, with constraints on the parameter space. Note that in population we have two (vector) equalities (7) and (13) indexed by parameters $(\widetilde{p}, \pi)$, satisfying the restrictions (8) and (9). Our parameter of interest is the counterfactual $\theta = \phi(\widetilde{p}, \pi; \mathfrak{p})$, that is, a low-dimensional function of the parameter $(\widetilde{p}, \pi)$. This feature makes recent procedures developed by Kaido, Molinari, and Stoye (2019) and Bugni, Canay, and Shi (2017) potentially useful tools for the problem treated in our paper. Both Kaido, Molinari, and Stoye (2019) and Bugni, Canay, and Shi (2017) propose highly versatile procedures that guarantee asymptotic uniform validity in a very general class of models. The nature of our proposal is somewhat different, in that we tailor it to address challenges inherent to empirical applications utilizing dynamic discrete choice models. The main hurdle, which can be potentially serious, in application of methods by Kaido, Molinari, and Stoye (2019) and Bugni, Canay, and Shi (2017) is computational. For example, take the classic model by Rust (1987) where we have a binary choice model, so $A = \widetilde{A} = 2$, with the dimension of the state space X (and $\widetilde{X}$) being 90. The dimension of $(\widetilde{p}, \pi)$ is then $(\widetilde{A} - 1)\widetilde{X} + AX = 270$. In fact, this is smaller than in typical empirical work; for instance, in a recent paper by Blundell, Gowrisankaran, and Langer (2020) the authors use $\widetilde{X} = X = 240$, making the relevant dimension 720. We conclude that in these empirical applications, the dimension of the original parameter space (before projection) tends to be prohibitively high for the practical application of this approach.

Alternatively, we can fix $\widetilde{p}$, but not π, and rewrite the system into a moment inequality form by eliminating π (i.e. solving for other variables). As noted in Kitamura and Stoye (2018), this amounts to transforming, in the language of discrete geometry, a $\mathcal{V}$-representation of a polytope to an $\mathcal{H}$-representation, and it is generally known to be expensive to compute, and impractical even for a moderately sized system.

Finally, one may try to eliminate both π and $\widetilde{p}$ from the system to get some form of moment inequalities; but this is even harder to implement, especially because of the nonlinear constraints that involve $\widetilde{p}$, and so it is not a practically feasible option either.

Remark 3. Obviously our procedure requires an appropriate choice of subsample size h_N. Fortunately, some practical guidelines have been proposed and used in the literature of set-identified models: see Bugni (2016) and Ciliberto and Tamer (2009). We adopt their recommendations in carrying out our subsamping procedure. Notably, Bugni (2016) provides theoretical rationale for the rate $h_N \asymp N^{2/3}$ and the efficacy of the choice $h_N = N^{2/3}$ has been confirmed in Bugni, Canay, and Shi (2017) through Monte Carlo simulations. In our set of Monte Carlo experiments, we have set $h_N = N^{2/3}$ and observed that it works well in our context. This, indeed, stands as our principal recommendation for practitioners. Alternatively, Ciliberto and Tamer (2009) propose a different option: $h_N = N/4$. Our simulation results confirm the suitability of this alternative in our setting as well, thus presenting a robust practical alternative or offering a valuable robustness check. The use of regularization through tuning parameters to overcome issues associated with nonregular features of statistical problems, such as the κ parameter in Andrews and Soares (2010), is standard, and it is always a good practice to verify robustness by trying different values of the tuning parameter. This is exactly what we report in our analysis.

6.1 Computational Algorithm for Inference

We now present our computational algorithm for our inference procedure, starting with a discussion on how we accurately approximate $\widehat{J}_N(\theta)$ and $\widehat{c}_{1-\alpha}(\theta)$. The algorithm's details are provided in Appendix C for reference.[25]

Computing $\widehat{J}_N(\theta_0)$ and $\widehat{c}_{1-\alpha}(\theta_0)$. To simplify, we focus on the scalar case, $\theta \in \mathbb{R}$. We first note that it may be difficult to solve the minimization problem (26) in the main sample and in all subsamples, especially when $\phi(\widetilde{p}, \pi; \mathfrak{p})$ is a complicated function, involving, e.g., the ergodic distribution of the state variables, as in our applied examples. That is because finding particular values for counterfactual CCP $\widetilde{p}$ and baseline flow payoff π that generate the *exact* (fixed) value θ_0 can be computationally burdensome. In fact, solving (26) *just once* is already extremely costly.

Our algorithm is designed to overcome this challenge, while simultaneously effectively handling high-dimensional problems. To do so, we take advantage of the relationship between the optimization problems (21)–(22) and (26). Specifically, and abstracting from sampling issues, consider the relaxed version of (21)–(22):

$$\theta^L(\epsilon) \equiv \min_{(\widetilde{p},\pi)\in\widetilde{\mathbf{P}}\times\mathbb{R}^{(A+1)X}} \phi\left(\widetilde{p}, \pi;\, \mathfrak{p}\right); \quad \theta^U(\epsilon) \equiv \max_{(\widetilde{p},\pi)\in\widetilde{\mathbf{P}}\times\mathbb{R}^{(A+1)X}} \phi\left(\widetilde{p}, \pi;\, \mathfrak{p}\right) \tag{30}$$

subject to

$$\begin{aligned}
\left\|\mathbf{M}(\mathfrak{p})\,\pi - b_{-J}(\mathfrak{p})\right\|_{\Omega} &\leq \epsilon, \\
R^{eq}\,\pi &= r^{eq}, \\
R^{iq}\,\pi &\leq r^{iq}, \\
(\widetilde{\mathbf{M}}(\mathfrak{p})\,\mathcal{H})\,\pi &= \widetilde{b}_{-J}\left(\widetilde{p}, \mathfrak{p}\right) - \widetilde{\mathbf{M}}(\mathfrak{p})\,g,
\end{aligned} \tag{31}$$

[25] Note that our algorithm is designed to address practical implementation challenges, which is distinct from the statistical process itself. Therefore, the approximation errors do not influence the outcomes in Theorem 1. In our Monte Carlo study, presented in Appendix D, we observe minimal approximation errors in the confidence sets, even when dealing with large state spaces.

where $\|.\|_\Omega$ is the matrix norm defined as $\|x\|_\Omega = x'\Omega x$ for $x \in \mathbb{R}^{AX}$, and $\epsilon \geq 0$.

The difference between the original problem (21)–(22) and its relaxed version (30)–(31) is the inequality constraint $\left\|\mathbf{M}(\mathfrak{p})\,\pi - b_{-J}(\mathfrak{p})\right\|_\Omega \leq \epsilon$. When $\epsilon = 0$, the problems coincide, and for any $\epsilon \geq 0$ it is evident that $\theta^L(\epsilon) \leq \theta^L(0) \equiv \theta^L \leq \theta^U \equiv \theta^U(0) \leq \theta^U(\epsilon)$. This implies that the identified set $[\theta^L, \theta^U] \equiv [\theta^L(0), \theta^U(0)]$ is contained in the interval $[\theta^L(\epsilon), \theta^U(\epsilon)]$ when $\epsilon > 0$.

Importantly, $J(\theta_0) = 0$ for all points θ_0 in the identified set $[\theta^L, \theta^U]$, and $J(\theta_0) \leq \epsilon$ for all points θ_0 in the wider interval $[\theta^L(\epsilon), \theta^U(\epsilon)]$ by construction. This means that for any θ_0 not in $[\theta^L, \theta^U]$ but within $[\theta^L(\epsilon), \theta^U(\epsilon)]$, with $\epsilon > 0$, we have that $0 < J(\theta_0) \leq \epsilon$. Consequently, by employing a sequence of ϵ values, $0 \equiv \epsilon_0 < \epsilon_1 < ... < \epsilon_k < ... < \epsilon_K \equiv \epsilon_{max}$, for some finite K, and solving the corresponding relaxed problem (30)–(31) for each ϵ_k, we obtain an increasing sequence of intervals:

$$[\theta^L(0), \theta^U(0)] \subseteq ... \subseteq [\theta^L(\epsilon_k), \theta^U(\epsilon_k)] \subseteq ... \subseteq [\theta^L(\epsilon_K), \theta^U(\epsilon_K)].$$

Simultaneously, we obtain a sequence of J values such that: (a) $J(\theta_0) = 0$ if $\theta_0 \in [\theta^L(0), \theta^U(0)]$, and (b) $\epsilon_{k-1} < J(\theta_0) \leq \epsilon_k$ if $\theta_0 \notin [\theta^L(\epsilon_{k-1}), \theta^U(\epsilon_{k-1})]$ and $\theta_0 \in [\theta^L(\epsilon_k), \theta^U(\epsilon_k)]$, for $k = 1, ..., K$. This enables us to obtain accurate approximations for $J(\theta_0)$ for any given θ_0 by using a fine grid of ϵ values and without having to solve the practically difficult problem (26) directly. Obtaining such approximations in the main sample and all subsamples is computationally cheaper than solving (26) repeatedly because (30)–(31) is a smooth well-behaved problem. In addition, because the relaxed problems based on ϵ_k and ϵ_{k+1} are very similar, the solution to the problem with ϵ_k provides an excellent initial value for the problem with ϵ_{k+1}, which helps reduce computational costs substantially.

In practice, we employ two step-functions, denoted as $\mathcal{J}^-(\theta)$ and $\mathcal{J}^+(\theta)$, to approximate $J(\theta)$ based on the grid points and the sequence of intervals $[\theta^L(\epsilon_k), \theta^U(\epsilon_k)]$. These approximations satisfy the following conditions: (i) $\mathcal{J}^-(\theta) = J(\theta) = \mathcal{J}^+(\theta) = \epsilon_0 = 0$ for $\theta \in [\theta^L, \theta^U]$, and (ii) $\mathcal{J}^-(\theta) \equiv \epsilon_{k-1} < J(\theta) \leq \epsilon_k \equiv \mathcal{J}^+(\theta)$ for $\theta \notin [\theta^L, \theta^U]$ and $k = 1, ..., K$. Essentially, $\mathcal{J}^-$ approximates J from below, while $\mathcal{J}^+$ approximates it from above. (See Figure C1 in Appendix C for an illustration.)

For a sufficiently fine grid, both approximations yield similar values. However, in the case of a coarse grid, to ensure conservative behavior in finite samples, we recommend approximating J in the main sample from below. Intuitively, this (weakly) reduces the value of the test statistic used to construct the confidence set, leading to over-coverage in finite samples, all else being constant. For the subsamples, we approximate J from above, as it (weakly) increases the critical value $\widehat{c}_{1-\alpha}$, which also leads to over-coverage in finite samples. Our Monte Carlo simulations confirm these intuitive results.

Step-by-Step Algorithm. Our computational approach consists of four main steps, explained in Algorithm 1 below. To facilitate exposition, we leave to Appendix C a detailed explanation of all auxiliary functions used in the algorithm.

The first step is to solve the $K+1$ relaxed problems (30)–(31) for the grid points $\boldsymbol{\epsilon} = \{\epsilon_k\}_{k=0}^{K}$ within the main sample, aimed at approximating (from below) the test statistic $N\widehat{J}_N(\theta)$ for any given θ. I.e., we approximate the *function* $\widehat{J}_N(\theta)$ in our first step. This is achieved through the use of two auxiliary functions, employed sequentially: (i) Approximations(.), which calculates the sequence of intervals, $[\theta^L(\epsilon_k), \theta^U(\epsilon_k)]$, with the corresponding solutions for π; and then (ii) $\mathcal{J}(.)$, which effectively approximates $\widehat{J}_N$ from below. Appendix C provides practical guidance on specifying the grid set $\boldsymbol{\epsilon}$ – defining values for ϵ_{max} and K for an equally spaced grid – and on selecting a set of initial values for the optimization problem (30)–(31) with $\epsilon_0 = 0$, referred to in line 3 of the algorithm. Of note, the randomly generated set of initial values that we recommend satisfy the model restrictions (7)–(9) by construction.

In the second step, we repeat the same approximation in each subsample, but now with three differences. First, we incorporate re-centering, as defined in (29), because it improves the finite sample behavior of the procedure. This is acomplished using the auxiliary function Recentering before the subsampling starts, as shown in line 6. Second, we rescale the grid set $\boldsymbol{\epsilon}$ with K equidistant points in the range from $\epsilon_0 = 0$ to $\frac{N}{h_N} \times \epsilon_{max}$. That is because failing to rescale the grid in the subsamples may lead to non-overlapping between the values that the test statistic $N\widehat{J}_N$ may take in the sample and the distribution of the test statistic in the subsamples $h_N\widehat{J}^*_{h_N}$, since $N \gg h_N$, which would invalidate the approximation to the critical value $\widehat{c}_{1-\alpha}$. We therefore adjust the inputs to the Approximations function accordingly, as presented in line 10 of the algorithm. Finally, we approximate the subsampling test statistic $\widehat{J}^*_{h_N}$ from above, as discussed previously.

In the third step, we obtain the critical value $\widehat{c}_{1-\alpha}(\theta)$ for any given θ from the subsampling approximations to $\widehat{J}^*_{h_N}$. In other words, in this step, we approximate the critical value *function*. Finally, in the last step, we construct the $1-\alpha$ confidence set $CS = [\theta^L_{ci}, \theta^U_{ci}]$ using the approximations to both functions $\widehat{J}_N(\theta)$ and $\widehat{c}_{1-\alpha}(\theta)$. At this point, it is trivial to determine the confidence set because the cost to evaluate the functions approximating $\widehat{J}_N(\theta)$ and $\widehat{c}_{1-\alpha}(\theta)$ at any given θ is negligible.

Algorithm 1 Subsampling Inference

```
Inputs: $\boldsymbol{\epsilon}$, $\mathfrak{p}$, $h_N$, $N$, $\alpha$, $S$    // $h_N = N^{2/3}$ recommended by Bugni (2016)
function CONFIDENCE_SET($\boldsymbol{\epsilon}$, $\mathfrak{p}$, $h_N$, $N$, $\alpha$, $S$)    // Compute the $1-\alpha$ confidence set for $\theta$
    $\boldsymbol{\pi}^{init} \leftarrow$ randomly generated set of initial values
    $(\widehat{\theta}^U(\cdot), \pi^U(\cdot), \widehat{\theta}^L(\cdot), \pi^L(\cdot)) \leftarrow$ APPROXIMATIONS$(\boldsymbol{\epsilon}, \boldsymbol{\pi}^{init}, \mathfrak{p}, 0, \text{main sample})$    // Compute sequence of intervals $[\theta^L(\epsilon_k), \theta^U(\epsilon_k)]$, with corresponding $\pi$'s, in the main sample
    $\widehat{J}(\cdot) \leftarrow \mathcal{J}(\widehat{\theta}^U(\cdot), \widehat{\theta}^L(\cdot), \boldsymbol{\epsilon}, \text{below})$    // Approximate $J$ from below in the main sample
    $b^{ctr}_{-J}(\cdot) \leftarrow$ RECENTERING$(\widehat{\theta}^U(\cdot), \pi^U(\cdot), \widehat{\theta}^L(\cdot), \pi^L(\cdot), \boldsymbol{\epsilon}, \mathfrak{p})$    // Recentering, used in the subsamples
    for $s = 1, \ldots, S$ do
        $\widehat{\mathfrak{p}}^*_s \leftarrow$ subsample with replacement of size $h_N$ from $\mathfrak{p}$.
        $\boldsymbol{\pi}^{init}_s \leftarrow$ randomly generated initial values
        $(\widehat{\theta}^{U*}_s(\cdot), \widehat{\theta}^{L*}_s(\cdot)) \leftarrow (\theta^U(\cdot), \theta^L(\cdot))$ from APPROXIMATIONS $\left(\frac{N}{h_N}\boldsymbol{\epsilon}, \boldsymbol{\pi}^{init}_s, \widehat{\mathfrak{p}}^*_s, b^{ctr}_{-J}(\cdot), \text{subsample}\right)$    // Compute sequence of intervals $[\theta^L(\epsilon_k), \theta^U(\epsilon_k)]$, with corresponding $\pi$'s, in each subsample
        $\widehat{J}^*_s(\cdot) \leftarrow \mathcal{J}(\widehat{\theta}^{U*}_s(\cdot), \widehat{\theta}^{L*}_s(\cdot), \boldsymbol{\epsilon}, \text{above})$    // Approximate $J$ from above in the subsamples
    end for
    $\widehat{c}_{1-\alpha}(\cdot) \leftarrow \text{quantile}\left(h_N \widehat{J}^*_s(\cdot), 1-\alpha\right)$    // Compute the $1-\alpha$ critical value function
    $\theta^L_{ci} \leftarrow \min\theta$ such that $N\widehat{J}(\theta) < \widehat{c}_{1-\alpha}(\theta)$    // Find the lower bound of the confidence set
    $\theta^U_{ci} \leftarrow \max\theta$ such that $N\widehat{J}(\theta) < \widehat{c}_{1-\alpha}(\theta)$    // Find the upper bound of the confidence set
    return $[\theta^L_{ci}, \theta^U_{ci}]$
end function
```

7 Empirical Application

In this section, we illustrate our approach in the context of a dynamic model of export behavior. To that end, we consider the setup of Das, Roberts, and Tybout (2007), henceforth 'DRT', who perform a horserace between different kinds of export subsidies. As the authors point out, industrial exporters are highly prized in developing countries for generating gains from trade, sustaining production and employment during domestic recessions, and facilitating the absorption of foreign technologies. As a consequence, exporters often receive governmental support. Yet, seemingly similar subsidies may generate different export responses in different industries and time periods, making it difficult for policy makers to know which type of support is optimal. To shed light on these issues, DRT develop a structural dynamic model of firm export decisions and simulate returns of different subsidies. Here, we adopt their specification and explore the identifying power of alternative model restrictions to assess which of their restrictions deliver their main findings.

Data. We use DRT's plant-level panel data from Colombian manufacturing industries and focus on the knitting mills industry. The dataset is composed of 64 knit fabric producers observed annually between 1981–1991; the sample has 704 plant-year observations. Like DRT, we focus on firms that operated continuously in the domestic market, given that they were responsible for most of the exports over this period. The share of exporting firms increased from 12% in 1981 to 18% by the end of the sample period, possibly a result of the 33% depreciation of Colombia's real exchange rate. This industry also depicts significant turnover: the average probability of re-entry into export markets is 61%. On average, export revenues of exporting firms are approximately 1.4 times the domestic revenues.

Model. DRT assume that export markets are monopolistically competitive; this leads to a specification similar to the firm entry/exit model presented in Section 5. In particular, every period t a firm i chooses whether to export or not, $a_{it} \in \mathcal{A} = \{0, 1\}$. The state variables are (i) the lagged decision ($k_{it} = a_{it-1}$), and (ii) exchange rates and demand/supply shocks in export markets (w_{it}). The exogenous shocks w_{it} follow (discretized) independent normal-AR(1) processes. The payoff function is given by equation (23) in Section 5. To point identify the model, DRT restrict to zero the payoffs of not exporting (i.e., both the outside value and the scrap value are set to zero). They also impose state-invariant entry and fixed costs, making their model overidentified. We relax these assumptions and instead explore the identifying power of Restrictions 1, 2, and 3 presented in the entry/exit model. In principle, scrap values may differ from zero because they may involve idleness costs (given that exiting is often temporary) or depreciation costs. Similarly, fixed costs and entry costs may depend on the aggregate states, as they involve finding trading partners, setting up distribution networks, maintaining labor and capital abroad, etc.[26] For ease of exposition, we leave the model details to Appendix I.

Counterfactuals and Outcomes of Interest. DRT focus on three counterfactual policies: (i) direct subsidies to plants' export revenues, such as a tax rebate that is proportional to foreign sales; (ii) subsidies to the cost of entering into exporting, such as grants for information or technology acquisition for export development; and (iii) payments designed to cover the annual fixed costs of operating in the export market. We follow DRT and consider a 2% export revenue subsidy, a 25% entry cost subsidy, and a 28% fixed cost subsidy.

The main outcome of interest is a benefit–cost ratio based on the long-run average annual gain in export revenues divided by the long-run average government subsidy expenditures. We denote the ratios for the revenue, fixed cost, and entry cost subsidies by θ_R, θ_F, and θ_E, respectively, and take $\theta = (\theta_R, \theta_F, \theta_E)$ – see Appendix I for explicit formulas for θ.

Evaluating ex-ante the impact of different model restrictions on θ is not trivial. Note first that while export revenues are observed in the data, the long-run *average* change in revenues depends on

[26] The payoff when not exporting (the outside option) may also be different from zero since it includes domestic profits. However, following DRT, we do not explore this possibility given the limitations in the data.

firms' decisions to export given the type of subsidy. This means that all numerators in θ depend on the counterfactual CCPs. Next, note that all denominators in θ equal the long-run average government expenditures, which depend on the fraction of firms exporting in the counterfactual steady-state; i.e., they all depend on $\widetilde{p}$ as well. In addition, θ_F and θ_E depend on the unknown parameters, fc and ec, respectively, since the government expenditures are direct functions of these costs. In the case of the entry cost subsidy, a further complication is that the (subsidized) entry cost is paid only when firms enter, implying that $\widetilde{p}$ affects the direct payments in each state (in addition to affecting the steady-state distribution). In short, θ depends on both $\widetilde{p}$ and π highly nonlinearly.

In terms of identification, the benefit-cost ratio of the revenue subsidy θ_R is point-identified. That is both because the averages in the numerator and denominator depend on observed revenues, and because $\widetilde{p}$ is identified (since it involves known changes to known quantities, i.e., the identified variable profits; see KSS), implying that the counterfactual steady-state distribution is also point-identified. The other two target objects, θ_F and θ_E, are partially identified both because (i) the counterfactual behavior $\widetilde{p}$ is not point-identified (as the entry subsidy in our example in Section 5), and (ii) the denominators in the benefit–cost ratios depend *directly* on model parameters that are partially identified (i.e., on fc and ec, respectively). In sum, both θ_F and θ_E involve ratios of set-identified objects.

Results. DRT find, under their imposed restrictions ($s = 0$, as well as ec, fc invariant over states), that revenue subsidies yield the highest return, followed by the fixed cost subsidies, and then by the entry cost subsidies; i.e. $\theta_R > \theta_F > \theta_E$. We explore the robustness of this finding under milder restrictions.

We implement our procedure as explained in Section 6 and in Appendices C and I.[27] Table 2 presents the benefit–cost ratios under Restrictions 1–3. The revenue subsidy generates an estimated benefit–cost ratio, θ_R, of approximately 15 pesos of revenue per unit cost. Its impact is statistically significant and economically large, and it is fairly consistent with the estimates in DRT. Because θ_R is point identified, it does not depend on any additional model restriction (other than the basic framework (7)).

We now discuss θ_F and θ_E, which are partially identified. Restriction 1 (i.e., $fc \geq 0$ and $ec \geq 0$) is not sufficiently informative here: the fixed cost subsidies ratio, θ_F, is between 8 and 30, and the entry cost subsidies ratio, θ_E, ranges from 4 to 24. These sets are wide because there are still many model parameter values that can rationalize observed behavior. The identified sets overlap and we cannot conclude which

[27] The transition process for exchange rates is taken from a long-time series as in DRT. Given the small sample size, we discretize the support of each exogenous state in three bins. We estimate CCPs using frequency estimators. To compute confidence intervals, we implement 1000 replications of a standard i.i.d. subsampling, resampling 16 firms over the sample time period, so that the size of each subsample is $h_N = 16 \approx N^{\frac{2}{3}}$. To minimize the quadratic distances in (27) and (28), we take a diagonal weighting matrix Ω with diagonal elements given by the square-root of the ergodic distribution of the state variable – thus, deviations on more visited states are considered more relevant and receive greater weights. Given that θ_R is known (ex ante) to be point identified, we use the plug-in estimator proposed by Kalouptsidi, Lima, and Souza-Rodrigues (2021) to estimate it, and 1000 standard i.i.d. bootstrap replications at the firm level to construct the confidence intervals for θ_R. To make our results comparable to DRT, we have also estimated the model parameters under their restrictions and obtained very similar results as theirs. See details in Appendix I.

Table 2: Export Revenue/Cost Ratio for Different Subsidies under Alternative Model Restrictions

	Restriction 1	Restrictions 1–2	Restrictions 1–3
		Revenue Subsidies	
Estimated Identified Set	15.13	15.13	15.13
90% Confidence Interval	(11.15, 18.90)	(11.15, 18.90)	(11.15, 18.90)
		Fixed Costs Subsidies	
Estimated Identified Set	[8.41, 30.82]	[11.10, 13.34]	[11.92, 12.60]
90% Confidence Interval	(7.33, 35.21)	(9.60, 14.64)	(9.42, 13.97)
		Entry Costs Subsidies	
Estimated Identified Set	[4.40, 24.04]	[7.85, 17.28]	[8.88, 9.36]
90% Confidence Interval	(3.42, 34.86)	(6.98, 23.89)	(7.14, 14.63)

Notes: This table shows the estimated sharp identified sets for the average gains in total export revenues divided by the average government subsidy expenditures, both averaged over states in the long-run. The top panel shows the gains of a 2% export revenue subsidy; the middle panel, the gains of a 28% fixed cost subsidy; and the bottom panel, the gains of a 25% entry cost subsidy. The (nonsingleton) identified sets are in brackets. The data set is composed of 704 plant-year observations in the Colombian knitting mills industry. The 90% confidence intervals are in parenthesis and were calculated based on 1000 bootstrap replications for the revenue subsidies, and 1000 subsample replications for both fixed and entry costs subsidies (with subsample sizes of $h_N = 16$). Restrictions 1, 2, and 3 are all specified in the main text (Section 5). See Online Appendix I for details.

policy generates the highest return.

Adding Restriction 2 increases the identification power substantially: the ratio for the fixed cost subsidies is now between 11 and 13. This identified set is highly informative and its upper bound is smaller than θ_R, suggesting that the revenue subsidy is more potent than the fixed cost subsidy. Intuitively, under Restriction 2, entry is profitable in the long-term, which imposes an upper limit on the values that fc can take. This upper limit, in turn, reduces the potential impact of the fixed cost subsidies.

Hence, under Restrictions 1 and 2 we can only confirm part of DRT's findings (that $\theta_R > \theta_F$). In contrast, there is substantial uncertainty regarding the benefit–cost ratio for the entry cost subsidy θ_E: its identified set is between 7.8 and 17, containing both the estimated θ_R and the identified set of θ_F.

Incorporating exclusion restrictions on scrap values (Restriction 3) narrows the identified set for θ_E substantially: the benefit-cost ratio now ranges from 8.9 to 9.4, which is highly informative. Hence, under Restrictions 1-3, we can confirm DRT's subsidy ranking ($\theta_R > \theta_F > \theta_E$): revenue subsidies generate the highest export revenues per unit cost, followed by fixed cost subsidies, and then by the entry cost subsidies. We thus conclude that, although this ranking can be obtained under milder restrictions than those imposed by DRT, it does seem to hinge on the assumption that scrap values do not depend on state variables. We now provide some intuition for this finding.

Why does the exclusion restriction on scrap values (Restriction 3) render entry cost subsidies less effective? Intuitively, under Restrictions 1–2, the state variables can induce a correlation between export revenues and scrap values. When that correlation is negative, entry subsidies would provide incentives for low-productivity firms (i.e., those with low export revenues) to enter and exit the export markets repeatedly, while high-productivity firms (with high export revenues) would stay longer in exporting. This makes the entry subsidy effective in terms of the benefit-cost ratio. When we impose the exclusion restriction on scrap values (Restriction 3), we eliminate that negative correlation and, as a result, the low-productivity firms stay more often in the export markets. This reduces the average gain in export revenues per peso of subsidy, pushing the upper bound on θ_E downwards, and making the entry subsidy worse than both the revenues and fixed costs subsidies. A similar result holds, but is less relevant in magnitude, for the lower bound on θ_E, reflecting possible positive correlations between export revenues and scrap values.

Of note, the uniformly valid confidence intervals indicate substantial sampling uncertainty, which is not surprising given the size of the data set.

8 Conclusion

In this paper, we explore how much one can learn about counterfactual outcomes in dynamic discrete choice models for a large and empirically relevant class of counterfactual experiments for which the level of flow payoffs may matter. We derive analytical properties of the identified sets under alternative model restrictions. We also develop an asymptotically uniformly valid inference approach based on subsampling, as well as novel and computationally tractable procedures that can handle high-dimensional problems – a prevalent issue in applied studies. The empirical implications of our results are illustrated by revisiting the study of Das, Roberts, and Tybout (2007) on exporting decisions and subsidies.

Our primary motivation is to offer a solution for practitioners that is applicable to a widely used class of empirical models. We hope our results can aid practitioners in assessing which empirical findings survive under minimal restrictions, in understanding the impact of commonly imposed restrictions on counterfactuals, and in including confidence sets around their counterfactual outcomes.

References

Adams, A. (2020): “Mutually Consistent Revealed Preference Demand Predictions,” *American Economic Journal: Microeconomics*, 12(1), 42–74.

Adao, R., A. Costinot, and D. Donaldson (2017): “Nonparametric Counterfactual Predictions in Neoclassical Models of International Trade,” *American Economic Review*, 107(3), 633–89.

Aguirregabiria, V. (2010): “Another Look at the Identification of Dynamic Discrete Decision Processes: An application to Retirement Behavior,” *Journal of Business & Economic Statistics*, 28(2), 201–218.

Aguirregabiria, V., and P. Mira (2002): “Swapping the Nested Fixed Point Algorithm: A Class of Estimators for Discrete Markov Decision Models,” *Econometrica*, 70(4), 1519–1543.

——— (2007): “Sequential Estimation of Dynamic Discrete Games,” *Econometrica*, 75(1), 1–53.

Aguirregabiria, V., and J. Suzuki (2014): “Identification and Counterfactuals in Dynamic Models of Market Entry and Exit,” *Quantitative Marketing and Economics*, 12(3), 267–304.

Andrews, D. W., and G. Soares (2010): “Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection,” *Econometrica*, 78(1), 119–157.

Arcidiacono, P., and R. A. Miller (2011): “Conditional Choice Probability Estimation of Dynamic Discrete Choice Models With Unobserved Heterogeneity,” *Econometrica*, 79(6), 1823–1867.

——— (2020): “Identifying Dynamic Discrete Choice Models off Short Panels,” *Journal of Econometrics*, 215(2), 473–485.

Bajari, P., C. L. Benkard, and J. Levin (2007): “Estimating Dynamic Models of Imperfect Competition,” *Econometrica*, 75(5), 1331–1370.

Bejara, M. (2020): “A Semi-structural Methodology for Policy Counterfactuals,” Discussion paper, MIT.

Berry, S., and G. Compiani (2020): “An Instrumental Variables Approach to Dynamic Models,” Discussion paper, Yale University.

Blundell, R., M. Browning, and I. Crawford (2008): “Best Nonparametric Bounds on Demand Responses,” *Econometrica*, (76), 1227–1262.

Blundell, R., D. Kristensen, and R. Matzkin (2014): “Bounding quantile demand functions using revealed preference inequalities,” *Journal of Econometrics*, 179(2), 112 – 127.

Blundell, W., G. Gowrisankaran, and A. Langer (2020): “Escalation of scrutiny: The gains from dynamic enforcement of environmental regulations,” *American Economic Review*, 110(8), 2558–2585.

Bugni, F. A. (2016): “Comparison of Inferential Methods in Partially Identified Models in Terms of Error in Coverage Probability,” *Econometric Theory*, 32(1), 187–242.

Bugni, F. A., I. A. Canay, and X. Shi (2017): “Inference for subvectors and other functions of partially identified parameters in moment inequality models,” *Quantitative Economics*, 8(1), 1–38.

Chernoff, H. (1954): “On the Distribution of the Likelihood Ratio,” *The Annals of Mathematical Statistics*, pp. 573–578.

Chiong, K. X., A. Galichon, and M. Shum (2016): "Duality in dynamic discrete-choice models," *Quantitative Economics*, 7(1), 83–115.

Christensen, T., and B. Connault (2021): "Counterfactual Sensitivity and Robustness," Discussion paper, New York University.

Ciliberto, F., and E. Tamer (2009): "Market structure and multiple equilibria in airline markets," *Econometrica*, 77(6), 1791–1828.

Das, S., M. J. Roberts, and J. R. Tybout (2007): "Market Entry Costs, Producer Heterogeneity, and Export Dynamics," *Econometrica*, 75(3), 837–873.

Dearing, A. (2019): "Pseudo-Value Functions and Closed-Form CCP Estimation of Dynamic Discrete Choice Models," Discussion paper, Ohio State University.

Dickstein, M. J., and E. Morales (2018): "What Do Exporters Know?," *The Quarterly Journal of Economics*, 133(4), 1753–1801.

Geyer, C. J. (1994): "On the asymptotics of constrained M-estimation," *The Annals of statistics*, pp. 1993–2010.

Hotz, V. J., and R. A. Miller (1993): "Conditional Choice Probabilities and the Estimation of Dynamic Models," *Review of Economic Studies*, 60(3), 497–529.

Hotz, V. J., R. A. Miller, S. Sanders, and J. Smith (1994): "A Simulation Estimator for Dynamic Models of Discrete Choice," *Review of Economic Studies*, 61(2), 265–89.

Hu, Y., and M. Shum (2012): "Nonparametric identification of dynamic models with unobserved state variables," *Journal of Econometrics*, 171(1), 32–44.

Ichimura, H., and C. Taber (2000): "Direct Estimation of Policy Impacts," *NBER Working Paper 254*.

——— (2002): "Semiparametric Reduced-Form Estimation of Tuition Subsidies," *American Economic Review*, 92(2), 286–292.

Kaido, H., F. Molinari, and J. Stoye (2019): "Confidence intervals for projections of partially identified parameters," *Econometrica*, 87(4), 1397–1432.

Kalouptsidi, M. (2014): "Time to build and fluctuations in bulk shipping," *The American Economic Review*, 104(2), 564–608.

Kalouptsidi, M., L. Lima, and E. Souza-Rodrigues (2021): "On Estimating Counterfactuals Directly in Dynamic Models," Discussion paper, University of Toronto.

Kalouptsidi, M., P. T. Scott, and E. Souza-Rodrigues (2017): “On the Non-identification of Counterfactuals in Dynamic Discrete Games,” *International Journal of Industrial Organization*, 50, 362–371.

Kalouptsidi, M., P. T. Scott, and E. A. Souza-Rodrigues (2021): “Identification of Counterfactuals in Dynamic Discrete Choice Models,” *Quantitative Economics*, 12(2), 351–403.

Kasahara, H., and K. Shimotsu (2009): “Nonparametric Identification of Finite Mixture Models of Dynamic Discrete Choices,” *Econometrica*, 77(1), pp. 135–175.

Kitamura, Y., and J. Stoye (2018): “Nonparametric Analysis of Random Utility Models,” *Econometrica*, 86(6), 1883–1909.

Kitamura, Y., and J. Stoye (2019): “Nonparametric Counterfactuals in Random Utility Models,” Discussion paper, Yale University.

Krantz, S. G., and H. R. Parks (2003): *The Implicit Function Theorem: History, Theory, and Applications*. Birkäuser Basel, 1 edn.

Magnac, T., and D. Thesmar (2002): “Identifying Dynamic Discrete Decision Processes,” *Econometrica*, 70(2), 801–816.

Manski, C. F. (2007): “Partial Identification of Counterfactual Choice Probabilities,” *International Economic Review*, 48, 1393–1410.

Marschak, J. (1953): “Economic Measurements for Policy and Prediction,” in *Cowles Commission Monograph 14: Studies in Econometric Methods*, ed. by W. C. Hood, and T. Koopmans. New York: Wiley.

McFadden, D. (1974): *Conditional logit analysis of qualitative choice behavior*chap. 4, pp. 105–142. Academic Press, New York.

Miller, R. (1984): “Job matching and occupational choice,” *Journal of Political Economy*, 92(6), 1086–1120.

Mogstad, M., A. Santos, and A. Torgovitsky (2018): “Using Instrumental Variables for Inference about Policy Relevant Treatment Parameters,” *Econometrica*, 86(5), 1589–1619.

Molinari, F. (2020): “Chapter 5 – Microeconometrics with Partial Identification,” in *Handbook of Econometrics, Volume 7A*, ed. by S. N. Durlauf, L. P. Hansen, J. J. Heckman, and R. L. Matzkin, vol. 7 of *Handbook of Econometrics*, pp. 355–486. Elsevier.

Morales, E., G. Sheu, and A. Zahler (2019): “Extended gravity,” *The Review of Economic Studies*, 86(6), 2668–2712.

Norets, A. (2011): “Semiparametric Identification of Dynamic Multinomial Choice Models,” Discussion paper, Brown University.

Norets, A., and X. Tang (2014): “Semiparametric Inference in dynamic binary choice models,” *The Review of Economic Studies*, 81(3), 1229–1262.

Pakes, A. (1986): “Patents as options: Some estimates of the value of holding European patent stocks,” *Econometrica*, 54(4), 755–784.

Pakes, A., M. Ostrovsky, and S. Berry (2007): “Simple estimators for the parameters of discrete dynamic games (with entry/exit examples),” *The RAND Journal of Economics*, 38(2), 373–399.

Pesendorfer, M., and P. Schmidt-Dengler (2008): “Asymptotic Least Squares Estimators for Dynamic Games,” *The Review of Economic Studies*, 75(3), 901–928.

Rockafellar, R. T., and R. J.-B. Wets (2009): *Variational analysis*, vol. 317. Springer Science & Business Media.

Romano, J. P., and A. M. Shaikh (2008): “Inference for identifiable parameters in partially identified econometric models,” *Journal of Statistical Planning and Inference*, 138(9), 2786–2807.

——— (2012): “On the uniform asymptotic validity of subsampling and the bootstrap,” *The Annals of Statistics*, 40(6), 2798–2822.

Rust, J. (1987): “Optimal Replacement of GMC Bus Engines: an Empirical Model of Harold Zurcher,” *Econometrica*, 55(5), 999–1033.

——— (1994): “Structural Estimation of Markov Decision Processes,” *Handbook of Econometrics 4*, 4, 3081–3143.

Shapiro, A. (1985): “Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints,” *Biometrika*, 72(1), 133–144.

Train, K. (2009): *Discrete Choice Methods with Simulation.* Cambridge UP.

Wolpin, K. (1984): “An estimable dynamic stochastic model of fertility and child mortality,” *Journal of Political Economy*, 92(5), 852–874.

Ziegler, G. M. (2012): *Lectures on Polytopes*, vol. 152. Springer Science & Business Media.

Online Appendix

A Proofs

A.1 Proof of Proposition 1

To simplify notation, we omit the dependencies of matrices and vectors on the observable $(p, F) \in \mathbf{P} \times \mathbf{F}$. First, we show that (i) the identified set $\widetilde{\mathbf{P}}^I$ is sharp. Then, we prove that (ii) $\widetilde{\mathbf{P}}^I$ is a smooth connected manifold with boundary, and with dimension in the interior given by $rank(\mathcal{C}_J\, \mathcal{Q}_J) \leq X - d$. Finally, we show that (iii) in the absence of equality restrictions (8), the dimension of the identified set simplifies to $rank(\mathcal{C}_J) \leq X$.

(i) The identified set $\widetilde{\mathbf{P}}^I$ defined in (14) is sharp by construction because equations (7), (8), and (9) contain all model restrictions, and equation (13) fully characterizes $\widetilde{p}$ as an (implicit) function of π (see the arguments in footnotes 10 and 12 in the main text, and the characterization of $\widetilde{\mathbf{P}}^I$ below).

(ii) Here, our proof proceeds in three steps. First, we reparameterize the model. Second, we construct a local function relating the counterfactual CCP and the (reparameterized) model parameters. Finally, we extend the local function to a global function whose image set characterizes the counterfactual identified set.

Reparameterization. Clearly, $\widetilde{\mathbf{P}}^I$ is empty whenever Π^I is empty, so we assume hereafter that Π^I is non-empty. Recall that the identified set is characterized by the equations (7), (8), (9), and (13). By combining (7) and (8), we get

$$(R^{eq}_{-J} M_{-J} + R^{eq}_J)\, \pi_J = r^{eq} - R^{eq}_{-J} b_{-J},$$

which is of the form:

$$Q^{eq} \pi_J = q^{eq}, \tag{A1}$$

where

$$Q^{eq} = R^{eq}_{-J} M_{-J} + R^{eq}_J \tag{A2}$$

is a $d \times X$ matrix, and $q^{eq} = r^{eq} - R^{eq}_{-J} b_{-J} \in \mathbb{R}^d$. Equation (A1) incorporates all equality restrictions on π, and expresses them in terms of the "free parameter" $\pi_J \in \mathbb{R}^X$.

We assume that $rank(Q^{eq}) = d$ and that the first d columns of Q^{eq} are independent.[28] We write (A1) as

$$Q^{eq} \pi_J = [Q_1\ Q_2] \begin{bmatrix} \pi^1_J \\ \kappa \end{bmatrix} = Q_1 \pi^1_J + Q_2 \kappa = q^{eq},$$

[28] In the more general case, the independent columns of Q^{eq} must be permuted to the front.

where Q_1 is $d \times d$ and non-singular. Then $\pi_J^1 = Q_1^{-1} q^{eq} - Q_1^{-1} Q_2 \kappa$. Therefore,

$$\pi_J = \begin{bmatrix} \pi_J^1 \\ \kappa \end{bmatrix} = \begin{bmatrix} -Q_1^{-1} Q_2 \\ I \end{bmatrix} \kappa + \begin{bmatrix} Q_1^{-1} q^{eq} \\ 0 \end{bmatrix} = \mathcal{Q}_J \kappa + r_Q, \tag{A3}$$

where, clearly,

$$\mathcal{Q}_J = \begin{bmatrix} -Q_1^{-1} Q_2 \\ I \end{bmatrix}. \tag{A4}$$

This gives a complete parameterization of all π_J in terms of the "free" $X - d$ parameters in the vector κ. Represent the elements of this set by $\pi_J(\kappa)$. Note that in the absence of the equality restrictions (8), we can just take $\pi_J = \kappa$.

Similarly, combine (7) and (9), to get

$$(R_{-J}^{iq} M_{-J} + R_J^{iq})\, \pi_J \leq r^{iq} - R_{-J}^{iq} b_{-J},$$

which is of the form:

$$Q^{iq} \pi_J \leq q^{iq},$$

where $Q^{iq} = R_{-J}^{iq} M_{-J} + R_J^{iq}$ is an $m \times X$ matrix, and $q^{iq} = r^{iq} - R_{-J}^{iq} b_{-J} \in \mathbb{R}^m$. Substituting π_J in the inequality above by $\pi_J(\kappa)$ defined in (A3) and rearranging, we get the m inequalities defined in terms of $\kappa \in \mathbb{R}^{X-d}$:

$$Q^{iq} \mathcal{Q}_J \kappa \leq q^{iq} - Q^{iq} r_Q. \tag{A5}$$

Define the set

$$\mathcal{K} = \left\{ \kappa \in \mathbb{R}^{X-d} : Q^{iq} \mathcal{Q}_J \kappa \leq q^{iq} - Q^{iq} r_Q \right\}. \tag{A6}$$

Clearly, $\mathcal{K}$ is a convex polyhedron. By construction, any vector $\pi = [\pi'_{-J}, \pi'_J]'$ such that $\pi_{-J} = M_{-J} \pi_J(\kappa) + b_{-J}$, with $\pi_J(\kappa)$ defined by (A3) for some $\kappa \in \mathcal{K}$ satisfies (7), (8), and (9). I.e., for any given $\kappa \in \mathcal{K}$, we can find one π satisfying all model restrictions.

Local Function. Combine (7) and (13) to obtain

$$\underbrace{[I, -\widetilde{M}_{-J}]}_{=\widetilde{\mathbf{M}}} \underbrace{\begin{bmatrix} \mathcal{H}_{1,-J} & \mathcal{H}_{1J} \\ \mathcal{H}_{2,-J} & \mathcal{H}_{2J} \end{bmatrix}}_{=\mathcal{H}} \underbrace{\begin{bmatrix} M_{-J} \pi_J + b_{-J} \\ \pi_J \end{bmatrix}}_{=\pi} = \widetilde{b}_{-J}(\widetilde{p}) - \widetilde{\mathbf{M}} g,$$

or,

$$\mathcal{C}_J \pi_J + \left(\mathcal{H}_{1,-J} - M_{-J} \mathcal{H}_{2,-J} \right) b_{-J} = \widetilde{b}_{-J}(\widetilde{p}) - g_{-J} + \widetilde{M}_{-J} g_J, \tag{A7}$$

where $\mathcal{C}_J$ is the $\widetilde{A}\widetilde{X} \times X$ matrix defined in equation (15).

Noting that $\widetilde{p}$ has to satisfy $\widetilde{X}$ restrictions as it is a collection of conditional probability vectors, let $\widetilde{p}^*$ denote a $\widetilde{A}\widetilde{X}$-vector of independent elements of $\widetilde{p}$, and denote the set of independent elements by $\widetilde{\mathbf{P}}^*$. Substitute (A3) into (A7), rearrange it, and define the function $\mathcal{F} : \mathbb{R}^{X-d} \times int(\widetilde{\mathbf{P}}^*) \to \mathbb{R}^{\widetilde{A}\widetilde{X}}$ given by

$$\mathcal{F}\left(\kappa, \widetilde{p}^*\right) = -\mathcal{C}_J \mathcal{Q}_J \kappa - \mathcal{C}_J r_Q - \left(\mathcal{H}_{1,-J} - M_{-J}\mathcal{H}_{2,-J}\right) b_{-J}(p) + \widetilde{b}_{-J}\left(\widetilde{p}^*\right) - g_{-J} + \widetilde{M}_{-J} g_J,$$

where $int(\widetilde{\mathbf{P}}^*)$ is the interior of $\widetilde{\mathbf{P}}^*$. Clearly, the model and counterfactual restrictions impose $\mathcal{F}\left(\kappa, \widetilde{p}^*\right) = 0$, for all $\kappa \in \mathcal{K}$.

The Jacobian of $\mathcal{F}$ is given by $\nabla\mathcal{F} = \left[\frac{\partial \mathcal{F}}{\partial \kappa}, \frac{\partial \mathcal{F}}{\partial \widetilde{p}^*}\right]$, with

$$\begin{aligned} \frac{\partial \mathcal{F}}{\partial \kappa} &= -\mathcal{C}_J \mathcal{Q}_J, \\ \frac{\partial \mathcal{F}}{\partial \widetilde{p}^*} &= \frac{\partial \widetilde{b}_{-J}}{\partial \widetilde{p}^*}. \end{aligned}$$

Because $\frac{\partial \widetilde{b}_{-J}}{\partial \widetilde{p}^*}$ is everywhere invertible (see KSS), the implicit function theorem applies. Specifically, for a point $(\kappa^0, \widetilde{p}^{*0}) \in \mathbb{R}^{X-d} \times int(\widetilde{\mathbf{P}}^*)$ satisfying $\mathcal{F}\left(\kappa^0, \widetilde{p}^{*0}\right) = 0$, there exist open sets $U \subseteq \mathbb{R}^{X-d}$ and $W \subseteq int(\widetilde{\mathbf{P}}^*)$ such that $\kappa^0 \in U$ and $\widetilde{p}^{*0} \in W$, and there exists a continuously differentiable function $\varphi : U \to W$ satisfying $\widetilde{p}^{*0} = \varphi(\kappa^0)$ and that

$$\mathcal{F}\left(\kappa, \varphi\left(\kappa\right)\right) = 0,$$

for all $\kappa \in U$. Furthermore,

$$\frac{\partial \varphi\left(\kappa\right)}{\partial \kappa} = -\left[\frac{\partial \mathcal{F}}{\partial \widetilde{p}^*}\right]^{-1} \frac{\partial \mathcal{F}}{\partial \kappa} = \left[\frac{\partial \widetilde{b}_{-J}}{\partial \widetilde{p}^*}\right]^{-1} \mathcal{C}_J \mathcal{Q}_J.$$

The rank of the matrix $\frac{\partial \varphi(\kappa)}{\partial \kappa}$ equals the rank of $\mathcal{C}_J \mathcal{Q}_J$ because $\frac{\partial \widetilde{b}_{-J}}{\partial \widetilde{p}^*}$ is invertible everywhere. Let $rank(\mathcal{C}_J \mathcal{Q}_J) = k$. By the Rank Theorem, the image set of φ is a differentiable k-dimensional manifold in $int(\widetilde{\mathbf{P}}^*)$ (see Theorem 3.5.1 in Krantz and Parks, 2003). Clearly, by restricting κ to the convex polyhedron $\mathcal{K}$, the image set of φ becomes a k-dimensional manifold with boundary.

Extension to a Global Function. We can construct a global function $\bar{\varphi}$ defined on the entire domain $\mathcal{K}$ based on the local function φ defined above. To do so, we need to show that the constructed $\bar{\varphi}$ is not a set-function on $\mathcal{K}$. I.e., if for any pair of points $(\kappa^0, \widetilde{p}^{*0})$ and $(\kappa^0, \widetilde{p}^{*1})$ with $\kappa^0 \in \mathcal{K}$ and $\widetilde{p}^{*0}, \widetilde{p}^{*1} \in int(\widetilde{\mathbf{P}}^*)$, if $\bar{\varphi}(\kappa^0) = \widetilde{p}^{*0}$ and $\bar{\varphi}(\kappa^0) = \widetilde{p}^{*1}$, then we must have $\widetilde{p}^{*0} = \widetilde{p}^{*1}$. Suppose by contradiction that there exist implicit functions φ^0 and φ^1 defined locally on the neighborhood of the points $(\kappa^0, \widetilde{p}^{*0})$ and $(\kappa^0, \widetilde{p}^{*1})$ such that $\widetilde{p}^{*0} = \varphi^0\left(\kappa^0\right)$ and $\widetilde{p}^{*1} = \varphi^1\left(\kappa^0\right)$, with $\widetilde{p}^{*0} \neq \widetilde{p}^{*1}$. Next, recall that for any point $\kappa^0 \in \mathcal{K}$, there exists only one vector of payoffs $\pi(\kappa^0) = [\pi'_{-J}(\kappa^0), \pi'_J(\kappa^0)]'$ satisfying all model restrictions: This vector is given by the elements $\pi_{-J}(\kappa^0) = M_{-J}\pi_J(\kappa^0)+b_{-J}$, and $\pi_J(\kappa^0)$ defined by (A3). This leads to the counterfactual payoff $\widetilde{\pi}(\kappa^0)$, which is given by the affine function $\widetilde{\pi}(\kappa^0) = \mathcal{H}\pi(\kappa^0) + g$. Finally, the counterfactual payoff $\widetilde{\pi}(\kappa^0)$ can generate just one conditional choice probability function in the counterfactual scenario (by the uniqueness of the solution of the Bellman equation). We therefore must have $\widetilde{p}^{*0} = \widetilde{p}^{*1}$ (as well as $\varphi^0 = \varphi^1 = \varphi$). The global function $\bar{\varphi}$ equals the local implicit functions everywhere.[29]

We conclude that the identified set $\widetilde{\mathbf{P}}^I$ is the image set of the global function $\bar{\varphi}$, defined on the domain $\mathcal{K}$. Consequently, $\widetilde{\mathbf{P}}^I$ is a manifold with boundary and with dimension in the interior given by the rank of $\mathcal{C}_J\mathcal{Q}_J$. Further, $\widetilde{\mathbf{P}}^I$ is connected because $\bar{\varphi}$ is a continuous function defined on the convex domain $\mathcal{K}$. Finally, we have $rank(\mathcal{C}_J\mathcal{Q}_J) \leq X - d$ because $rank(\mathcal{C}_J) \leq \min\{\widetilde{A}\widetilde{X}, X\}$ and $rank(\mathcal{Q}_J) = X - d$.

(ii) In the absence of model restrictions, we take $\pi_J = \kappa$, or equivalently, $\mathcal{Q}_J = I$ and $r_Q = 0$. Following the same argument as in part (i), we conclude that $\widetilde{\mathbf{P}}^I$ is a connected manifold with boundary and with dimension in the interior given by the rank of $\mathcal{C}_J$.

A.2 Proof of Proposition 2

Fix $(p, F) \in \mathbf{P} \times \mathbf{F}$ and omit it from our notation. First we show that the rows of the matrix $\mathcal{C}_J$ corresponding to indices in $\mathbb{L}'$ are zero and therefore do not contribute to the rank of $\mathcal{C}_J$. Then we derive the rank of $\mathcal{C}_J$ based on the rows corrresponding to indices in $\mathbb{L}$.

We start by stacking (2) for all $a \neq J$ to obtain

$$\pi_{-J} = M_{-J}\pi_J + b_{-J}, \tag{A8}$$

where M_{-J} stacks M_a for all $a \neq J$. Consider the index $l \in \mathbb{L}'$. Take the l^{th} entry of $\widetilde{\pi}_{-J}$, denoted by $\widetilde{\pi}_{-J}(l)$. Let M_l denote the l row of M_{-J} Then, the counterfactual version of equation (A8) for the l entry gives

$$\widetilde{\pi}_{-J}(l) = M_l\widetilde{\pi}_J + b_l\left(\widetilde{p}\right) = M_l\pi_J + b_l\left(\widetilde{p}\right), \tag{A9}$$

where b_l corresponds to the l^{th} entry of the vector b_{-J}. Moreover, $\widetilde{\pi}_{-J}(l) = \pi_{-J}(l)$ because $l \in \mathbb{L}'$, while again from (A8),

$$\widetilde{\pi}_{-J}(l) = \pi_{-J}(l) = M_l\pi_J + b_l(p). \tag{A10}$$

(Here, we opted for emphasizing the dependency of b_l on p explicitly.) Equating the right hand side of (A9) and (A10) gives $b_l\left(\widetilde{p}\right) = b_l\left(p\right)$. Since this holds for any π_J, equation (A7) implies that all rows of $\mathcal{C}_J$ located in $\mathbb{L}'$ are equal to zero.

Next consider the set of indices $\mathbb{L}$. The factual and counterfactual versions of (A8) restricted on $\mathbb{L}$

[29] While different κ's can generate the same $\widetilde{p}^*$ (because the function φ is not one-to-one, which is at the heart of the identification problem in dynamic discrete choice models), a single κ cannot generate more than one $\widetilde{p}^*$.

gives

$$\pi_{-J}(\mathbb{L}) = M_{-J}(\mathbb{L})\pi_J + b_{-J}\left(p\right)(\mathbb{L}), \tag{A11}$$

$$\widetilde{\pi}_{-J}(\mathbb{L}) = M_{-J}(\mathbb{L})\widetilde{\pi}_J + b_{-J}\left(\widetilde{p}\right)(\mathbb{L}) = M_{-J}(\mathbb{L})\pi_J + b_{-J}\left(\widetilde{p}\right)(\mathbb{L}), \tag{A12}$$

where $\pi_{-J}(\mathbb{L})$, $\widetilde{\pi}_{-J}(\mathbb{L})$, $M_{-J}(\mathbb{L})$, $b_{-J}\left(p\right)(\mathbb{L})$, and $b_{-J}\left(\widetilde{p}\right)(\mathbb{L})$ select all the entries $\pi_{-J}(l)$, $\widetilde{\pi}_{-J}(l)$, $M_{-J}(l)$, $b_l(p)$, and $b_l\left(\widetilde{p}\right)$ for $l \in \mathbb{L}$; and the right most equality is due to the invariance of J, $\widetilde{\pi}_J = \pi_J$. The definition of the counterfactual transformation, $\widetilde{\pi} = \mathcal{H}\pi$, implies

$$\widetilde{\pi}(\mathbb{L}) = \mathcal{H}(\mathbb{L})\pi(\mathbb{L}) + \mathcal{H}(\mathbb{L}')\pi(\mathbb{L}') = \mathcal{H}(\mathbb{L})\pi(\mathbb{L}),$$

where the second equality holds because, by assumption, $\mathcal{H}(\mathbb{L}') = 0$. Next, notice that, because $\widetilde{\pi}_J = \pi_J$, we have that $\widetilde{\pi}(\mathbb{L}) = \widetilde{\pi}_{-J}(\mathbb{L})$ and $\pi(\mathbb{L}) = \pi_{-J}(\mathbb{L})$, so that the equality above becomes $\widetilde{\pi}_{-J}(\mathbb{L}) = \mathcal{H}(\mathbb{L})\pi_{-J}(\mathbb{L})$. We now insert π_J above using (A11) to obtain,

$$\widetilde{\pi}_{-J}(\mathbb{L}) = \mathcal{H}(\mathbb{L})M_{-J}(\mathbb{L})\pi_J + \mathcal{H}(\mathbb{L})b_{-J}(p)(\mathbb{L}).$$

Next, using (A12),

$$b_l\left(\widetilde{p}\right)(\mathbb{L}) = (\mathcal{H}(\mathbb{L}) - I)M_{-J}(\mathbb{L})\,\pi_J + \mathcal{H}(\mathbb{L})b_{-J}(p)(\mathbb{L}).$$

Comparison to (A7) demonstrates that the nonzero rows of $\mathcal{C}_J$ are given by $[\mathcal{H}(\mathbb{L}) - I]M_{-J}(\mathbb{L})$. Therefore, the rank of $\mathcal{C}_J$ satisfies

$$rank(\mathcal{C}_J) \leq \#\left\{\text{eigenvalues of } \mathcal{H}(\mathbb{L}) \text{ different from } 1\right\} \leq L,$$

because $M_{-J}(\mathbb{L})$ is full rank. The first inequality becomes an equality when $\mathcal{H}(\mathbb{L})$ is diagonalizable.

A.3 Proof of Proposition 3

Fix $(p, F) \in \mathbf{P} \times \mathbf{F}$ and omit it from our notation. First, note that the specification (16) can be equivalently described by linear restrictions (8) (see Appendix E). Then, the proof follows from Proposition 1. Here we show the connection $\mathcal{Q}_J = z_J$. To that end, take equation (A7) and replace π_J by its parametric specification defined in (16), so that

$$\mathcal{C}_J\, z_J\, \gamma + \mathcal{C}_J \delta_J + \left(\mathcal{H}_{1,-J} - M_{-J}\mathcal{H}_{2,-J}\right) b_{-J} = \widetilde{b}_{-J}\left(\widetilde{p}\right) - g_{-J} + \widetilde{M}_{-J} g_J.$$

Clearly, the dimension of the identified set $\widetilde{\mathbf{P}}^I$ is given by $rank(\mathcal{C}_J z_J) \leq \eta_\gamma$.[30]

A.4 Proof of Proposition 4

Although this is a special case of Proposition 1, it is simpler to prove it directly. Fix $(p, F) \in \mathbf{P} \times \mathbf{F}$ and omit it from our notation. Consider our base equation (2) for both the baseline and counterfactual scenarios, for all $a \in \mathcal{A}$:

$$\pi_a = M_a \pi_J + b_a(p) \tag{A13}$$

$$\widetilde{\pi}_a = M_a \widetilde{\pi}_J + b_a(\widetilde{p}). \tag{A14}$$

Subtract (A13) from (A14):

$$\widetilde{\pi}_a - \pi_a = M_a(\widetilde{\pi}_J - \pi_J) + b_a(\widetilde{p}) - b_a(p). \tag{A15}$$

Next, use the fact that $\pi_a = z_a \gamma + \delta_a$ and $\widetilde{\pi}_a = z_a \widetilde{\gamma} + \delta_a$ for all actions:

$$z_a(\widetilde{\gamma} - \gamma) = M_a z_J(\widetilde{\gamma} - \gamma) + b_a(\widetilde{p}) - b_a(p).$$

Or

$$(z_a - M_a z_J)(\widetilde{\gamma} - \gamma) = b_a(\widetilde{p}) - b_a(p). \tag{A16}$$

Now, notice that $\widetilde{\gamma}(\mathbb{L}') - \gamma(\mathbb{L}') = 0$ and that

$$\widetilde{\gamma}(\mathbb{L}) - \gamma(\mathbb{L}) = (\mathcal{D} - I)\,\gamma(\mathbb{L}) + g(\mathbb{L}),$$

where I is a conformable identity matrix. Thus, (A16) becomes

$$(z_a(\mathbb{L}) - M_a z_J(\mathbb{L}))\,(\mathcal{D} - I)\,\gamma(\mathbb{L}) + (z_a(\mathbb{L}) - M_a z_J(\mathbb{L}))\,g(\mathbb{L}) = b_a(\widetilde{p}) - b_a(p).$$

By stacking over $a \neq J$, we get

$$\mathbf{M}\,Z(\mathbb{L})\,(\mathcal{D} - I)\,\gamma(\mathbb{L}) + \mathbf{M}\,Z(\mathbb{L})\,g(\mathbb{L}) = b_{-J}(\widetilde{p}) - b_{-J}(p), \tag{A17}$$

which corresponds to equation (A7). The result follows.

A.5 Proof of Proposition 5

Fix $(p, F) \in \mathbf{P} \times \mathbf{F}$ and omit it from our notation. The identified set $\mathbf{\Theta}^I$ defined in (20) is sharp by construction. Following the proof of Proposition 1, we can construct payoff vectors satisfying all model

[30] Equivalently, one can reparametrize π_J in (A3) in terms of the η_γ "free parameters" γ, instead of the $X - d$ "free parameters" κ. The rest of the proof follows.

restrictions, denoted by $\pi(\kappa)$, and obtain the counterfactual CCP from the function $\widetilde{p}^* = \bar{\varphi}(\kappa)$, where $\bar{\varphi}$ is continuously differentiable, $\kappa \in \mathcal{K}$, and $\mathcal{K}$ is defined in (A6). We have therefore

$$\theta = \phi(\widetilde{p},\, \pi) = \phi(\bar{\varphi}(\kappa),\, \pi(\kappa)) = \bar{\phi}(\kappa),$$

When the function ϕ is continuous, so is the function $\bar{\phi}$ because $\bar{\varphi}(\kappa)$ and $\pi(\kappa)$ are both continuous. Clearly, $\mathbf{\Theta}^I$ equals the image set of the function $\bar{\phi}$ defined on the domain $\mathcal{K}$. The image set is connected because $\mathcal{K}$ is convex, and it becomes compact when $\mathcal{K}$ is compact (which happens when Π^I is bounded, see the proof of Proposition 1). Furthermore, when θ is a scalar, the connected set $\mathbf{\Theta}^I$ becomes an interval.

A.6 Proof of Theorem 1

First we introduce conditions used in the statement of Theorem 1. Formally, the (populational) minimization problem (26) projects $b_{-J}(\mathfrak{p})$ on the manifold $S(\mathfrak{p},\theta)$ under the weighted norm $\|x\|_\Omega = x'\Omega x$, for $x \in \mathbb{R}^{AX}$, where

$$S(\mathfrak{p},\theta) := \{\mathbf{M}(\mathfrak{p})\pi, \pi \in \mathbb{R}^{(A+1)X} : \mathcal{R}(\theta,\pi,\widetilde{p};\, \mathfrak{p}) = 0, \text{ and } R^{iq}\pi \leq r^{iq} \text{ hold for some } \widetilde{p} \in \widetilde{\mathbf{P}}\}. \tag{A18}$$

The value of $J(\theta)$ is the squared length of the projection residual vector.

It is useful to impose a mild requirement on $S(\mathfrak{p},\theta)$ in terms of its local geometric property. To this end, we introduce the notion of tangent cone:

Definition 1. *For a (possibly non-convex) set $A \subset \mathbb{R}^d$, the tangent cone of A at $x \in A$, henceforth denoted by $T_A(x)$, is given by*

$$T_A(x) := \limsup_{\tau \downarrow 0} \tau^{-1}(A \ominus x),$$

where $\ominus$ denotes the usual Minkowski difference.

See, e.g., Section 6A of Rockafellar and Wets (2009) for a discussion on the role of a tangent cone and other related concepts.

Condition 1. *(i) For every $(\mathfrak{p},\theta_0) \in \mathcal{P}$, the tangent cone $T_{S(\mathfrak{p},\theta_0)}(x)$ of $S(\mathfrak{p},\theta_0)$ is convex at each $x \in \mathbb{R}^{AX} \in S(\mathfrak{p},\theta_0)$. (ii) The tuning parameter sequence h_N is chosen such that $h_N \to \infty$ and $h_N/N \to 0$ as $N \to \infty$. (iii) Both ϕ and h_F are continuously differentiable functions.*

Condition 1 is weak and reasonable. Condition 1(ii) is a restriction on the subsample size, our tuning parameter. Condition 1(iii) imposes a standard smoothness restriction on counterfactuals. Condition 1(i) is also a very weak restriction, and many of existing tools for asymptotically uniform inference for set-identified models cease to remain valid when this is violated. More importantly, it holds for many commonly encountered applications. For example, this trivially holds for the counterfactual considered in our Monte Carlo because θ is the long-run average probability of staying in the market and so it does not directly depend on π. Next, consider the three counterfactual objects in the DRT empirical application (see Appendix I). Once again, Condition 1(i) holds trivially for the first specification, θ_R, regarding the benefit-cost ratio of the export revenue subsidy, for the same reason as above. With the second example concerning a fixed cost subsidy, θ_F, note that the set $S(\mathfrak{p}, \widetilde{p}, \theta) := \{\mathbf{M}(\mathfrak{p})\pi, \pi \in \mathbb{R}^{(A+1)X} : \mathcal{R}(\theta, \pi, \widetilde{p};\, \mathfrak{p}) = 0\}$ in this case is a linear manifold determined by the ergodic distribution of the state variables in the counterfactual scenario, $\widetilde{f}^{*}$, which is continuous in $\widetilde{p}$. Therefore, in this example $S^{*}(\mathfrak{p}, \theta) := \cup_{\widetilde{p} \in \widetilde{\mathbf{P}}} S(\mathfrak{p}, \widetilde{p}, \theta)$ satisfies Condition 1(i). As the set $\{\mathbf{M}(\mathfrak{p})\pi, \pi \in \mathbb{R}^{(A+1)X} : R^{iq}\pi \leq r^{iq}\}$ is a convex polytope, we see that Condition 1(i) holds for θ_F. A similar argument shows that it holds for the third counterfactual concerning an entry cost subsidy, θ_E, as well.

Now we prove Theorem 1. Consider a sequence $\{(\mathfrak{p}_N, \theta_N) \in \mathcal{P}, N \in \mathbb{N}\}$. Recall that p and F are determined by $\mathfrak{p}$. Let $(p_N, F_N) := (p(\mathfrak{p}_N), F(\mathfrak{p}_N))$. In what follows we use symbols such as S_N, $\widehat{S}_N$, $\bar{\Pi}_N$, $(\bar{V}_N, V_N, v)$, $(\bar{W}_N, W_N, w)$, B, μ_N, $\binom{\eta}{\zeta}$ and Σ (and their appropriate subsample counterparts with an asterisk symbol * in superscript) while omitting their dependence on θ_N to ease the notational burden in the proof.

Let $S_N := S(\mathfrak{p}_N, \theta_N)$ and $\widehat{S}_N := S(\widehat{\mathfrak{p}}_N, \theta_N)$, where $S(\mathfrak{p}, \theta)$ is defined in (A18). Then writing $\|x\|^2_\Omega := x'\Omega x$ for $x \in \mathbb{R}^{AX}$,

$$\begin{aligned} N\widehat{J}_N(\theta_N) &= \min_{x \in \widehat{S}_N} N\|b_{-J}(\widehat{\mathfrak{p}}_N) - x\|^2_{\widehat{\Omega}_N} \\ &= \min_{x \in \widehat{S}_N} \|\sqrt{N}[b_{-J}(\widehat{\mathfrak{p}}_N) - b_{-J}(\mathfrak{p}_N)] - \sqrt{N}[x - b_{-J}(\mathfrak{p}_N)]\|^2_{\widehat{\Omega}_N} \\ &= \min_{\xi \in \sqrt{N}(\widehat{S}_N \ominus b_{-J}(\mathfrak{p}_N))} \|\sqrt{N}[b_{-J}(\widehat{\mathfrak{p}}_N) - b_{-J}(\mathfrak{p}_N)] - \xi\,\|^2_{\widehat{\Omega}_N}, \end{aligned} \tag{A19}$$

where $\ominus$ denotes the usual Minkowski difference, and for $c \in \mathbb{R}_{++}$ and a set $A \in \mathbb{R}^d$, we let cA denote the set A dilated by the factor c, that is, $\{cx : x \in A\}$.

To show the theorem, it suffices to consider sequences $\mathfrak{p}_N, N \in \mathbb{N}$ such that

(i) $\inf_{x \in \text{bdy}(S_N)} \|b_{-J}(\mathfrak{p}_N) - x\|_\Omega = O(1/\sqrt{N})$, where $\text{bdy}(S_N)$ is the boundary of S_N, and

(ii) Each sequence $\{\mathfrak{p}_N, N = 1, 2, ...\}$ converges.

Suppose $\mathfrak{p}_N, N \in \mathbb{N}$ satisfies (i) and (ii). The restrictions imposed on $\mathcal{P}$ guarantee that along the sequence $\mathfrak{p}_N$ it holds that

$$\sqrt{N}[b_{-J}(\widehat{\mathfrak{p}}_N) - b_{-J}(\mathfrak{p}_N)] \xrightarrow{d} \eta,$$

where η is a zero mean Gaussian vector. In what follows we also use the following notation: for finite sets $V, W \subseteq \mathbb{R}^d$ we let $\text{conv}(V)$ and $\text{cone}(W)$ denote the convex hull of V and the cone spanned by W, respectively; then the Minkowski sum $\text{conv}(V) \oplus \text{cone}(W)$ is a polyhedron. We approximate the last term in equation (A19) following Chernoff (1954). Under Condition 1 we have:

$$N\widehat{J}_N(\theta_N) \overset{d}{=} \min_{\xi\in\bar{\Pi}_N} \|\eta-\xi\|_\Omega^2 + o_p(1), \tag{A20}$$

where $\bar{\Pi}_N = \mathrm{conv}(\bar{V}_N) \oplus \mathrm{cone}(\bar{W}_N)$ is a random polyhedron, with $\bar{V}_N = V_N + v$, $V_N \in \mathbb{R}^{AX\times m}$, $\bar{W}_N = W_N + w$, $W_N \in \mathbb{R}^{AX\times n}$, and v and w are $\mathbb{R}^{AX\times m}$-valued and $\mathbb{R}^{AX\times n}$-valued zero-mean Gaussian random matrices, respectively, for some $m, n \in \mathbb{N}$. Put loosely, the polyhedron $\bar{\Pi}_N$ is determined through vertices and rays, parameterized by $\bar{V}_N$ and rays $\bar{W}_N$, respectively. This corresponds to the so-called $\mathcal{V}$-representation (see, for example, Theorem 1.2 in Ziegler (2012)) of the polyhedron $\bar{\Pi}_N$, which is represented as a sum of a convex hull of the AX-vectors $\bar{V}_N$ and conical combination of the AX-vectors $\bar{W}_N$, which approximates $\widehat{S}_N := S(\widehat{\mathfrak{p}}_N, \theta_N)$.

Note that the estimation uncertainty in $\widehat{S}_N$ makes the polyhedron $\bar{\Pi}_N$ that appears in the asymptotic approximation (A20) random. Also define a (deterministic) sequence of polyhedra $\Pi_N = \mathrm{conv}(V_N) \oplus \mathrm{cone}(W_N)$. By the representation theorem for polyhedra (see, for example, Theorem 1.2 in Ziegler (2012)) we can write

$$\Pi_N = \{\xi : B\xi \le \mu_N\} \text{ for some } B \in \mathbb{R}^{\ell\times AX},$$

where $\mu_N \ge 0$ for all N and $\mu_N = O(1)$.

Recalling that each transition matrix $F_a, a \in \mathcal{A}$, depends on $\mathfrak{p}_N$ (as so does F), write

$$\det(M_a(\mathfrak{p}_N)) = \det\left((1-\beta F_a(\mathfrak{p}_N))(1-\beta F_J(\mathfrak{p}_N))^{-1}\right) = \frac{\det(I-\beta F_a(\mathfrak{p}_N))}{\det(I-\beta F_J(\mathfrak{p}_N))}.$$

Let $\{\lambda_a^i(\mathfrak{p}_N)\}$ and $\{\lambda_J^i(\mathfrak{p}_N)\}$ be the eigenvalues of $F_a(\mathfrak{p}_N)$ and $F_J(\mathfrak{p}_N)$, then

$$\begin{aligned}\det(M_a(\mathfrak{p}_N)) &= \frac{\det(\beta^{-1}I - F_a(\mathfrak{p}_N))}{\det(\beta^{-1}I - F_J(\mathfrak{p}_N))}\\ &= \frac{\prod_{i=1}^X(\beta^{-1}-\lambda_a^i(\mathfrak{p}_N))}{\prod_{i=1}^X(\beta^{-1}-\lambda_J^i(\mathfrak{p}_N))}\\ &> c, \qquad \text{for every } a\in\mathcal{A} \text{ and every } N\in\mathbb{N} \end{aligned} \tag{A21}$$

holds for some $c > 0$ that does not depend on N as β is fixed in the unit interval $(0, 1)$ and $\{\lambda_a^i(\mathfrak{p}_N)\}$ and $\{\lambda_J^i(\mathfrak{p}_N)\}$ are inside the unit circle for every N.

Note that the approximation (A20) holds for any sequence $\{V_N', W_N'\}_{N\in\mathbb{N}}$ such that $V_N' = V_N + o(1)$ and $W_N' = W_N + o(1)$, and with Condition 1 and (A21) we can choose $\{V_N, W_N\}_{N\in\mathbb{N}}$ such that the matrix B above does not depend on N. Then we have an alternative representation for the random polyhedron $\bar{\Pi}_N$ as well: for some positive definite matrix Σ it holds that

$$\bar{\Pi}_N = \{\xi : B\xi \le \mu_N + \zeta\},$$

where the vector $\binom{\eta}{\zeta} \sim \mathrm{N}(0, \Sigma)$. In sum, we have

$$N\widehat{J}_N(\theta_N) \overset{d}{=} \min_{\xi : B\xi \leq \mu_N + \zeta} \|\eta - \xi\|_\Omega^2 + o_p(1). \tag{A22}$$

Next we turn to the subsample statistic $\widehat{J}^*_{h_N}(\theta_N)$. To show the uniform validity of subsampling we can instead analyze the asymptotic behavior of the statistic $\widehat{J}_{h_N}$, the $\widehat{J}$-statistic calculated from a random sample of size h_N, drawn according to $\mathfrak{p}_N$ (Romano and Shaikh, 2012). That is, we now study the limiting behavior of the CDF $G_{h_N}(x, \mathfrak{p}_N), N = 1, 2, ...,$ where $G_\ell(x, \mathfrak{p}) := \Pr_{\mathfrak{p}}\{\ell \widehat{J}_\ell(\theta_N) \leq x\}$ for $\ell \in \mathbb{N}$. Then proceeding as before, along the sequence $\mathfrak{p}_N$ we have

$$h_N \widehat{J}_{h_N}(\theta_N) \quad \overset{d}{=} \quad \min_{\xi \in \bar{\Pi}^*_{h_N,N}} \|\eta^* - \xi\|_\Omega^2 + o_p(1), \tag{A23}$$

where $\bar{\Pi}^*_{h_N,N} = \mathrm{conv}(\bar{V}^*_{h_N,N}) \oplus \mathrm{cone}(\bar{W}^*_{h_N,N})$, with $\bar{V}^*_{h_N,N} = V^*_{h_N,N} + v^*$, $V^*_{h_N,N} \in \mathbb{R}^{AX \times m}$, $\bar{W}^*_{h_N,N} = W^*_{h_N,N} + w^*$, $W^*_{h_N,N} \in \mathbb{R}^{AX \times m}$, and η^*, v^* and w^* are zero-mean Gaussian random elements taking values in $\mathbb{R}^{AX}$, $\mathbb{R}^{AX \times m}$ and $\mathbb{R}^{AX \times n}$ with $(\eta^*, v^*, w^*) \overset{d}{=} (\eta, v, w)$. Define $\Pi^*_{h_N,N} = \mathrm{conv}(V^*_{h_N,N}) \oplus \mathrm{cone}(W^*_{h_N,N})$ and observe that it has a half-space based representation $\Pi^*_{h_N,N} = \{\xi : B\xi \leq \sqrt{\frac{h_N}{N}}\mu_N\}$. We now have

$$\bar{\Pi}^*_N = \left\{\xi : B\xi \leq \sqrt{\frac{h_N}{N}}\mu_N + \zeta^*\right\}.$$

Recall that $\mu_N = \mathrm{O}(1)$, and moreover, we have $\binom{\eta^*}{\zeta^*} \sim \mathrm{N}(0, \Sigma)$. Therefore

$$h_N \widehat{J}_{h_N}(\theta_N) \xrightarrow{d} \min_{\xi : B\xi \leq \zeta} \|\eta - \xi\|_\Omega^2. \tag{A24}$$

In sum, for every sequence $\mathfrak{p}_N, N \in \mathbb{N}$ satisfying conditions (i) and (ii) above, by (A22) and (A24) and noting $\mu_N \geq 0$ for every N, we have

$$\limsup_{N \to \infty} \sup_x (G_{h_N}(x, \mathfrak{p}_N) - G_N(x, \mathfrak{p}_N)) \leq 0.$$

We can now invoke Theorem 2.1 in Romano and Shaikh (2012) to conclude.

B A Proposed Stochastic Search Algorithm for the Identified Set, When Analytic Gradients are not Available

As explained in the main text, in our experience, standard solvers are highly efficient in solving (21)–(22) when the researcher can provide the gradient of ϕ. However, when numerical (or analytical) gradients are costly to evaluate in practice, standard solvers can be slow in converging to the optimum. For such cases, we propose a stochastic algorithm that exploits the structure of the problem and combines the strengths of alternative stochastic search procedures, as we explain below.

Our proposed algorithm builds upon a couple of observations. First, while a search over π to maximize ϕ is feasible, it is computationally costly especially when it is expensive to calculate the gradient of ϕ numerically. In high-dimensional problems, this may become intractable. This procedure searches over the admissible values that π can take, and, for each candidate, finds the corresponding counterfactual CCP by solving the nonlinear equation (13), and then it evaluates ϕ – and its (numerical) derivative, to obtain updated directions for π – until reaching the maximum value for θ. Although finding admissible values for π is not difficult in high-dimensional problems (as it only depends on linear constraints), and solving the nonlinear equation (13) once is not computationally costly (as standard quasi-Newton methods can be used to find $\widetilde{p}$), solving (13) too many times and calculating the gradient of ϕ numerically can be demanding. Unless the econometrician imposes a sufficient number of assumptions to make π effectively a low dimensional vector (e.g., 3-dimensional or smaller), this method takes a long time to converge, as it requires too many evaluations before we can increase θ substantially in the direction of its maximum.

Second, it is possible to perform a search over $\widetilde{p}$, instead of π, to calculate θ^U. For any given $\widetilde{p}$, existence of a π satisfying linear constraints is computationally cheap (for example, existence can be easily checked as a solution to a linear programming problem). If there is no such π satisfying all restrictions, we discard $\widetilde{p}$, since it does not belong to the identified set $\widetilde{\mathbf{P}}^I$. If there exists *some* π satisfying the restrictions, we keep $\widetilde{p}$, and compute the corresponding θ. This approach may be particularly useful when ϕ is not a direct function of π, in which case it is not necessary to find a particular π to calculate θ – existence of *some* π suffices. The difficulty here is that, while an exhaustive grid search over $\widetilde{p}$ can be used to find the maximum θ^U, grid search is unfeasible for empirically-relevant high-dimensional problems. An alternative would be to perform a stochastic search (to find good directions for $\widetilde{p}$).[31] Yet, and more importantly, the random search must be performed on the $\widetilde{A}\widetilde{X}$–space $\widetilde{\mathbf{P}}$, while the identified set $\widetilde{\mathbf{P}}^I$ can be of much smaller dimension: $X - d$, or smaller (depending on the rank of $\mathcal{C}_J\mathcal{Q}_J$; see Proposition 1). In other words, $\widetilde{\mathbf{P}}^I$ may be a "thin" set in $\widetilde{\mathbf{P}}$. The combination of a "thin" set with an unknown shape (recall that $\widetilde{p}$ is a nonlinear function of π – see equation (13)) makes it difficult to find points within that set randomly. Further, it is easy for perturbation methods to "exit" the set, increasing the costs of finding the maximum θ. Note that searching over π to maximize θ does not suffer from this problem because finding admissible values (and updated directions) for π are computationally easier.

These trade-offs led us to consider an algorithm that exploits the structure of the problem and combines the strengths of these alternative search procedures. Intuitively, we move in the "$\widetilde{p}$-world" (to avoid solving

[31] For instance, one possibility is to perturb $\widetilde{p}$ completely randomly ($\widetilde{p} + \varepsilon$) and check whether the perturbed vector lies in the identified set $\widetilde{\mathbf{P}}^I$ (or within a tolerance level) – where checking this amounts to checking existence of π satisfying linear restrictions, as mentioned above. We then keep the perturbed $\widetilde{p}'s$ that deliver large values for θ (and perturb them further), and discard those with small values of θ. We iterate until θ cannot be increased any longer. (This is similar to genetic algorithm, or to stochastic search methods more generally.)

the nonlinear equation (13) repeatedly), but we keep a close eye on the "π-world" (to keep track of the model restrictions and search in relevant directions). Searching in relevant directions without solving (13) and computing the numerical gradient of ϕ in every step improves substantially how fast θ moves on each iteration to the maximum.

B.1 The Algorithm

We now present our proposed algorithm. To guarantee that $\widetilde{p}$ are positive and add up to one, we work with the transformation

$$\widetilde{\delta} = \ln \widetilde{p}_{-J} - \ln \widetilde{p}_J,$$

where $\ln \widetilde{p}_a$ is the $\widetilde{X} \times 1$ vector with elements $\ln \widetilde{p}_a(x)$, for all $x \in \widetilde{\mathcal{X}}$, and $\ln \widetilde{p}_{-J}$ stacks $\ln \widetilde{p}_a$ for all $a \neq J$. The functions of $\widetilde{p}$, namely $\widetilde{b}_{-J}$ and ϕ, are adjusted accordingly. To simplify the notation, we drop the subscript of the function $\widetilde{b}_{-J}$, as well as the argument p of the function ϕ. The algorithm proceeds as follows (we discuss the most important steps below):

```
The Proposed Stochastic Search Algorithm:

 1. Initialize $k = 0 \in \mathbb{N}$.
    Set $\pi^k$ satisfying $R^{eq}\pi^k = r^{eq}$ and $R^{iq}\pi^k \leq r^{iq}$.  Find $\widetilde{\delta}^k$ by solving (13) with $\pi^k$.
    Calculate $\theta^k = \phi(\widetilde{\delta}^k, \pi^k)$.

 2. Increment $k$.

 3. Set (perturbed) direction $\Delta\pi^k$.  Given $\Delta\pi^k$, set direction for $\widetilde{\delta}^k$,
        $$\Delta\widetilde{\delta}^k = \left(\frac{\partial \widetilde{b}}{\partial \widetilde{\delta}}\right)^{-1} \widetilde{\mathbf{M}}\, \mathcal{H}\, \Delta\pi^k,$$
    where $\left(\frac{\partial \widetilde{b}}{\partial \widetilde{\delta}}\right)$ is the derivative of $\widetilde{b}$ with respect to $\widetilde{\delta}$ evaluated at $\widetilde{\delta}^k$.

 4. Solve for $\alpha \in \mathbb{R}$:
        $$\alpha^* = \underset{\alpha}{\operatorname{argmax}} \;\; \phi(\,\widetilde{\delta}^k + \alpha\Delta\widetilde{\delta}^k, \pi^k + \alpha\Delta\pi^k\,),$$
    subject to the constraints (22), allowing (13) to be violated at most by a
    tolerance level, $\text{tol} > 0$.

 5. Set $\widetilde{\delta}^* = \widetilde{\delta}^k + \alpha^*\Delta\widetilde{\delta}^k$ and $\pi^* = \pi^k + \alpha^*\Delta\pi^k$.

 6. Update $\widetilde{\delta}^k$:
        $$\widetilde{\delta}^{k+1} = \widetilde{\delta}^* - \left(\frac{\partial \widetilde{b}^*}{\partial \widetilde{\delta}}\right)^{-1} \left(\widetilde{b}^* - \widetilde{\mathbf{M}}g - \widetilde{\mathbf{M}}\mathcal{H}\pi^*\right),$$
```

```
   where $\left(\frac{\partial \widetilde{b}^*}{\partial \widetilde{\delta}}\right)$ is the derivative of $\widetilde{b}$ with respect to $\widetilde{\delta}$ evaluated at $\widetilde{\delta}^*$, and define
   $\widetilde{b}^* = \widetilde{b}(\widetilde{\delta}^*)$.  Set $\pi^{k+1} = \pi^*$.

7. Calculate $\theta^{k+1} = \phi(\widetilde{\delta}^{k+1}, \pi^{k+1})$.
   If $\left\|\theta^{k+1} - \theta^k\right\| \leq \epsilon$ go to 8; otherwise go to 2.

8. Set $\pi = \pi^{k+1}$.  Solve (13) exactly for $\pi$, and get $\widetilde{\delta}$.  Return $\theta^U = \phi(\widetilde{\delta}, \pi)$.
```

We now discuss the rationale for each step. In Subsection B.2, we provide further details for the implementation of each step, as well as a discussion about the overall cost of the algorithm.

Step 1. The first step requires finding a π that satisfies the model restrictions (8) and (9) so that we obtain an initial $\widetilde{p}$ (or $\widetilde{\delta}$) that lies inside the (potentially "thin") set $\widetilde{\mathbf{P}}^I$ by construction, and a corresponding θ in the identified set $\mathbf{\Theta}^I$. Such initial π can be obtained as any solution to the following quadratic programming problem

$$\min_{\pi} \; \left(R^{eq}\pi - r^{eq}\right)' \left(R^{eq}\pi - r^{eq}\right) + (R^{iq}\pi - r^{iq})'_+(R^{iq}\pi - r^{iq})_+, \tag{B1}$$

where $(x)_+ = \max\{x, 0\}$. Another option is to start with a few points and project them into the identified set for π, which can also be done easily. Of note, if the minimum of (B1) is strictly greater than zero, then there is no π that satisfies all the constraints. Given π, we can solve (13) numerically using some quasi-Newton method.

Step 3. After we have our starting point π (and corresponding $\widetilde{\delta}$), we need to obtain an updated direction $\Delta\pi$ (and $\Delta\widetilde{\delta}$). Overall, the idea of first providing a direction and only then optimizing (as we do here) is a standard way to solve complex optimization problems. Ideally, we would use the gradient of ϕ, but calculating this gradient can be expensive in some cases, as mentioned previously. An alternative is to either get a completely random direction for $\Delta\pi$ (e.g., $\Delta\pi = \eta$, where η is a random vector drawn from, say, a multivariate standard normal distribution), or a random direction weighted by states that are more important (e.g., in terms of the ergodic distribution of the state variables).[32]

It is also important to not let an updated point get too close to the boundary of the inequality constraints (9). We follow the insights of interior-point methods to help the algorithm not get stuck early on a boundary. Specifically, we add a term to $\Delta\pi$ that moves it way from the most binding ones.

[32] In practice, to weight the random direction η by states that are more important in terms of the steady-state distribution, we draw η from a normal distribution with zero mean and a diagonal variance-covariance matrix with a diagonal that equals the probabilities of the state variables under the ergodic distribution. The ergodic distribution is based on the latest updated $\widetilde{p}$.

Formally,

$$\Delta\pi = \eta - \lambda \left(\frac{1}{r^{iq} - R^{iq}\pi} \right)' R^{iq},$$

where $\lambda = \frac{\lambda_0}{N}$, with $\lambda_0 > 0$ and $N =$ the number of iterations; and $\left(\frac{1}{r^{iq} - R^{iq}\pi} \right)$ denotes the $m \times 1$ vector with the reciprocal elements of the vector $r^{iq} - R^{iq}\pi$ (recall that m is the number of inequality restrictions, so that R^{iq} is $m \times (A+1)X$ and r^{iq} is $m \times 1$). The adjustment term $\lambda \left(\frac{1}{r^{iq} - R^{iq}\pi} \right)' R^{iq}$ is a common way to handle inequality constraints. This is a simple implementation of an interior-point method.[33]

We link the direction $\Delta\widetilde{\delta}$ with $\Delta\pi$ based on equation (13). We do so because completely random directions on $\widetilde{p}$ (or more precisely, on $\widetilde{\delta}$) will likely push $\widetilde{p}$ outside of the "thin" set $\widetilde{\mathbf{P}}^I$. The direction $\Delta\widetilde{\delta}$ is obtained by differentiating the inverse function $\widetilde{b}^{-1}$ with respect to π in the direction $\Delta\pi$.

Step 4. Given $\Delta\widetilde{\delta}$, we now find how far in that direction we should go without moving away too much from the identified set $\widetilde{\mathbf{P}}^I$. To that end, we allow for small violations in equation (13) when searching for α^*. Specifically, we replace the restriction (13) by $\left\| \widetilde{b} - \widetilde{\mathbf{M}}g - \widetilde{\mathbf{M}}\mathcal{H}\pi \right\| \leq \text{tol}$, where $\|.\|$ is some matrix norm and tol > 0 is a tolerance level. Here, the optimization is one-dimensional (line-search). We use a simple golden rule search, but even more crude approaches work.

Step 5. We now update both $\widetilde{\delta}$ and π in their respective directions $\alpha^*\Delta\widetilde{\delta}$ and $\alpha^*\Delta\pi_J$, where α^* is obtained in step 4.

Step 6. This step is important because at the end of step 5 it is common that the intermediary $\widetilde{\delta}^*$ violates the nonlinear system (13) by the maximum tolerance tol. So this step insures that we move $\widetilde{\delta}$ back to the set that violates (13) by strictly less than tol. Not doing so would constrain the directions that $\Delta\widetilde{\delta}$ can move in the next iteration and slow down the algorithm considerably.

Step 7. The $\epsilon > 0$ in step 7 specifies the tolerance for convergence. We focus on convergence on θ because verifying a "derivative equals zero" condition for convergence is difficult given the high-dimensionality of the problem and the complexity of computing derivatives of ϕ (analytically or numerically).

Step 8. After convergence, we solve the nonlinear system (13) exactly to guarantee that $\widetilde{p}$ lies in the identified set $\widetilde{\mathbf{P}}^I$, and so that the computed θ^U belongs to $\mathbf{\Theta}^I$.

[33] Intuitively, to maximize $f(x)$ subject to $g(x) \leq 0$, an interior-point method can make use of the logarithmic "barrier function" $B(x, \lambda) = f(x) - \lambda \sum_{i=1}^{n} \log(g_i(x))$, where n is the dimension of g. The gradient of B is $\frac{\partial f}{\partial x} - \lambda \sum_{i=1}^{n} \frac{1}{g(x)} \frac{\partial g}{\partial x}$. The idea is that when some element $g_i(x)$ is close to zero for some trial x, the barrier function "explodes" to minus infinity, so that the algorithm does not get stuck on a boundary. However, because the solution may indeed lie on the boundary, it is necessary to allow for the possibility that $g_i(x) = 0$ at the optimum. To do so, λ must converge to zero as the number of iterations grows larger. In the present case, we take $\lambda = \frac{\lambda_0}{N} \to 0$ (as $N \to \infty$). The term $\left(\frac{1}{r^{iq} - R^{iq}\pi} \right)' R^{iq}$ is the derivative of the sum of the logs of $(r^{iq} - R^{iq}\pi)$ with respect to π (i.e., the derivative of $\lambda \sum log(r^{iq} - R^{iq}\pi)$, where the summation runs from 1 to m).

One of the main computational cost of this algorithm is to calculate the inverse matrix $\left(\frac{\partial \widetilde{b}}{\partial \widetilde{\delta}}\right)^{-1}$, used in steps 3 and 6. In the next subsection we discuss conditions under which calculating $\left(\frac{\partial \widetilde{b}}{\partial \widetilde{\delta}}\right)^{-1}$ is not as costly.

B.2 Further Comments on Implementation

We now comment on the computational costs of the algorithm.

1. The matrix M_a equals $(I - \beta F_a)(I - \beta F_J)^{-1}$, which involves the inversion of a $X \times X$ matrix. The computational cost of inverting a matrix is of the order of $O(X^3)$ in general, but there are ways to reduce this cost. When action J is renewal or terminal, the matrix simplifies to $M_a = I + \beta (F_J - F_a)$, for all $a \in \mathcal{A}$, which can be calculated fast since it involves no matrix inversion (see footnote 7 in Online Appendix G). When there are no terminal or renewal actions, computing M_a requires calculating the inverse of $(I - \beta F_J)$. As F_J is a transition matrix, we can approximate that inverse based on the geometric series:
$$(I - \beta F_J)^{-1} = \sum_{\tau=0}^{\infty} \beta^{\tau} F_J^{\tau}.$$
By truncating the series, we can reduce the computational cost and obtain a reasonable approximation (more on that below). Note that both M_a and $\widetilde{M}_a$ can be precomputed, so they do not add costs to the iterative procedure.

2. When we find the direction $\Delta\widetilde{\delta}$ implied by $\Delta\pi$ we need to solve the linear system
$$\Delta\widetilde{\delta} = \left(\frac{\partial \widetilde{b}}{\partial \widetilde{\delta}}\right)^{-1} \widetilde{\mathbf{M}}\,\mathcal{H}\Delta\pi.$$
Usually this would cost $O(A^3X^3)$. However, we can take advantage of the structure of the function $\widetilde{b}$. Recall that $\widetilde{b}_a(\widetilde{p}) = \widetilde{M}_a\widetilde{\psi}_J(\widetilde{p}) - \widetilde{\psi}_a(\widetilde{p})$. Let $\widetilde{\psi}_{-J}$ stack $\widetilde{\psi}_a$ for all actions $a \neq J$. For expositional convenience, consider the case of three actions with reference action $J = 3$:
$$\widetilde{b} = \widetilde{M}_{-J}\widetilde{\psi}_J - \widetilde{\psi}_{-J} = \widetilde{\delta} - \left(\begin{bmatrix}\mathbf{I}\\ \mathbf{I}\end{bmatrix} - \widetilde{M}_{-J}\right)\log\left(1 + \sum_{j=1}^{J-1}\exp(\widetilde{\delta}_j)\right),$$
where $\mathbf{I}$ is the identity matrix. So
$$\frac{\partial \widetilde{b}}{\partial \widetilde{\delta}} = \mathbf{I} - \left(\begin{bmatrix}\mathbf{I}\\ \mathbf{I}\end{bmatrix} - \widetilde{M}_{-J}\right)\begin{bmatrix}\widetilde{P}_1 & \widetilde{P}_2\end{bmatrix},$$
where $\widetilde{P}_j$ is an $X \times X$ diagonal matrix with $\widetilde{p}_j$ as its entries.

Now we need its inverse

$$\left(\frac{\partial \widetilde{b}}{\partial \widetilde{\delta}}\right)^{-1} = \mathbf{I} + \left(\begin{bmatrix}\mathbf{I}\\ \mathbf{I}\end{bmatrix} - \widetilde{M}_{-J}\right)\left(\mathbf{I} - \begin{bmatrix}\widetilde{P}_1 & \widetilde{P}_2\end{bmatrix}\left(\begin{bmatrix}\mathbf{I}\\ \mathbf{I}\end{bmatrix} - \widetilde{M}_{-J}\right)\right)^{-1}\begin{bmatrix}\widetilde{P}_1 & \widetilde{P}_2\end{bmatrix}$$

$$= \mathbf{I} + \left(\begin{bmatrix}\mathbf{I}\\ \mathbf{I}\end{bmatrix} - \widetilde{M}_{-J}\right)\left(\mathbf{I} - \widetilde{P}_1 - \widetilde{P}_2 + \widetilde{P}_1\widetilde{M}_1 + \widetilde{P}_2\widetilde{M}_2\right)^{-1}\begin{bmatrix}\widetilde{P}_1 & \widetilde{P}_2\end{bmatrix},$$

where the first equality follows from the Woodbury formula $(\mathbf{I} - DB)^{-1} = \mathbf{I} + D(\mathbf{I} - BD)^{-1}B$.

Next, notice that $\widetilde{P}_J = \mathbf{I} - \widetilde{P}_1 - \widetilde{P}_2$ and that $\widetilde{M}_j = (\mathbf{I} - \beta\widetilde{F}_j)(\mathbf{I} - \beta\widetilde{F}_J)^{-1}$. Therefore,

$$\left(\frac{\partial \widetilde{b}}{\partial \widetilde{\delta}}\right)^{-1} = \mathbf{I} + \left(\begin{bmatrix}\mathbf{I}\\ \mathbf{I}\end{bmatrix} - \widetilde{M}_{-J}\right)(\mathbf{I} - \beta\widetilde{F}_J)\left(\widetilde{P}_J(\mathbf{I} - \beta\widetilde{F}_J) + \widetilde{P}_1(\mathbf{I} - \beta\widetilde{F}_1) + \widetilde{P}_2(\mathbf{I} - \beta\widetilde{F}_2)\right)^{-1}\begin{bmatrix}\widetilde{P}_1 & \widetilde{P}_2\end{bmatrix}$$

$$= \mathbf{I} + \beta\begin{bmatrix}\widetilde{F}_1 - \widetilde{F}_J\\ \widetilde{F}_2 - \widetilde{F}_J\end{bmatrix}\left(\mathbf{I} - \beta\left(\widetilde{P}_J\widetilde{F}_J + \widetilde{P}_1\widetilde{F}_1 + \widetilde{P}_2\widetilde{F}_2\right)\right)^{-1}\begin{bmatrix}\widetilde{P}_1 & \widetilde{P}_2\end{bmatrix}.$$

This reduces the cost to $O(X^3)$ because the matrix

$$\left(\mathbf{I} - \beta\left(\widetilde{P}_J\widetilde{F}_J + \widetilde{P}_1\widetilde{F}_1 + \widetilde{P}_2\widetilde{F}_2\right)\right)^{-1}$$

has dimension $X \times X$.

But we can improve on that by noticing that for a given vector v,

$$\left(\mathbf{I} - \beta\left(\widetilde{P}_J\widetilde{F}_J + \widetilde{P}_1\widetilde{F}_1 + \widetilde{P}_2\widetilde{F}_2\right)\right)^{-1} v = \sum_{\tau=0}^{\infty}\beta^{\tau}\left(\widetilde{P}_J\widetilde{F}_J + \widetilde{P}_1\widetilde{F}_1 + \widetilde{P}_2\widetilde{F}_2\right)^{\tau} v.$$

Because $\widetilde{P}_J\widetilde{F}_J + \widetilde{P}_1\widetilde{F}_1 + \widetilde{P}_2\widetilde{F}_2$ is a transition matrix, we know that

$$\left(\widetilde{P}_J\widetilde{F}_J + \widetilde{P}_1\widetilde{F}_1 + \widetilde{P}_2\widetilde{F}_2\right)^{\tau} v \to v^*$$

for some v^* as τ goes to infinity.[34] Therefore, we can approximate

$$\left(\mathbf{I} - \beta\left(\widetilde{P}_J\widetilde{F}_J + \widetilde{P}_1\widetilde{F}_1 + \widetilde{P}_2\widetilde{F}_2\right)\right)^{-1} v \approx \sum_{\tau=0}^{K-1} \beta^\tau \left(\widetilde{P}_J\widetilde{F}_J + \widetilde{P}_1\widetilde{F}_1 + \widetilde{P}_2\widetilde{F}_2\right)^\tau v + \frac{\beta^K}{1-\beta}\left(\widetilde{P}_J\widetilde{F}_J + \widetilde{P}_1\widetilde{F}_1 + \widetilde{P}_2\widetilde{F}_2\right)^K v,$$

which can be computed in $O(KX^2)$ operations. K can be taken small because we only need a reasonable approximation (and, as long as the exogenous states are not too persistent, it should mix fast).

C Inference Procedure

We now provide the details of our computational algorithm for inference, followed by our recommendations for the choice of initial values for the optimization problems and the tuning parameters. Then, we discuss the computational costs of our proposed approach.

C.1 Computational Algorithm

In this section, we focus on the auxiliary functions used in Algorithm 1 presented in Section 6.1 of the main paper. Recall that our approach consists of four main steps: (i) we approximate $\widehat{J}_N$ from below in the main sample; (ii) we approximate $\widehat{J}^*_{h_N}$ from above in each subsample; (iii) we obtain the $1-\alpha$ critical values $\widehat{c}_{1-\alpha}$; and, finally, (iv) we construct the $1-\alpha$ confidence set $CS = [\theta^L_{ci}, \theta^U_{ci}]$ using the approximations to both functions $\widehat{J}_N(\theta)$ and $\widehat{c}_{1-\alpha}(\theta)$.

In the first step of Algorithm 1, we make use of the function Approximations(.), which calculates the sequence of intervals, $[\theta^L(\epsilon_k), \theta^U(\epsilon_k)]$, with the corresponding solutions for π. This function is explained in Algorithm 2 below. It distinguishes whether the calculations are done in the main sample (`type = sample`) or in the subsamples (`type = subsample`), and it proceeds in four stages. First, for each proposed vector of initial value π^{init}_q, it solves the optimization problems (21)–(22) in the main data, regardless of the `type` selected – as shown in lines 4–5. To solve these problems, it uses the auxiliary function Optimization(.), presented in Algorithm 3, which solves the relaxed problems (30)–(31) for any given ϵ.[35] Then, if the approximations are done in the subsamples (`type = subsample`), the routine repeats the optimizations in each subsample, with recentering – incorporated via the input $b^{ctr}_{-J}(.)$, evaluated at either θ^U_q or θ^L_q, obtained from the solution calculated in the main sample. Next, it selects the best outcomes for θ^U, θ^L, π^U, and π^L based on the different initial values used, as presented in line 13.

[34] Under the ℓ_1 norm, this convergence is a contraction and the contraction coefficient is known as Dobrushin ergodic coefficient.

[35] Although not explicit in Algorithm 3, the Optimization function also receives as an input the definition of the function of interest ϕ, as well as the analytical gradient of ϕ if provided by the researcher (otherwise, it calculates the numerical derivative, as usual). One can also provide the optimization method, e.g., the Newton method.

Importantly, all the calculations up to this point are for $\epsilon_0 = 0$. Finally, in the fourth stage, we solve the relaxed problem (30)–(31) in the grid set $\boldsymbol{\epsilon} = \{\epsilon_k\}_{k=1}^{K}$, using as the initial value for each ϵ_k, $k = 1, ..., K$, the solution obtained from the previous grid point ϵ_{k-1}, in sequence. This last step is performed either in the main data or in the subsamples depending on the input `type` selected.

After the sequence of intervals $[\theta^L(\epsilon_k), \theta^U(\epsilon_k)]$, with their corresponding π's, are obtained on the grid $\boldsymbol{\epsilon}$, Algorithm 1 approximates $\widehat{J}_N$ from below using the auxiliary function $\mathcal{J}(.)$. The calculation of this auxiliary function $\mathcal{J}(.)$ is presented at the middle panel of Algorithm 3. It takes the set of intervals $[\theta^L(\epsilon_k), \theta^U(\epsilon_k)]$, the grid points $\boldsymbol{\epsilon}$, and the type of approximation (either from below or from above) as inputs and returns a piecewise constant function $\mathcal{J}^{\pm} \colon \Theta \to \mathbb{R}$.

In the third step of the main Algorithm 1, the same approximations done in the main sample are repeated in all subsamples (but now from above). There, as noted before, we employ recentering to improve finite sample performance. To do so, we use the auxiliary function Recentering(.), presented at the bottom panel of Algorithm 3. This function takes the set of intervals $[\theta^L(\epsilon_k), \theta^U(\epsilon_k)]$, together with their corresponding π's, the grid points $\boldsymbol{\epsilon}$, and the main data $\mathfrak{p}$ as inputs and returns a piecewise constant function from Θ to $\mathbb{R}^{X(A-1)}$. Recentering is a function of θ because the problem (28), upon which the recentering is defined (see equation (29)), depends explicitly on the value of θ under consideration.

Algorithm 2 Auxiliary Function: APPROXIMATIONS

```
function APPROXIMATIONS($\boldsymbol{\epsilon}$, $\boldsymbol{\pi}^{init}$, $\mathfrak{p}$,$b^{ctr}_{-J}(\cdot)$, type)
    $b_{-J} \leftarrow -\mathbf{M}(\mathfrak{p})\boldsymbol{\psi}(\mathfrak{p})$    // Compute vector $b_{-J}$, based on equation (4)
    for $q = 1, \dots, \#\boldsymbol{\pi}^{init}$ do
        $(\theta^U_q, \pi^U_q) \leftarrow \text{OPTIMIZATION}(0, \pi^{init}_q, b_{-J}(\mathfrak{p}), \mathfrak{p}, \max)$ // Solve (21)–(22) for $\max \phi$ in the sample
        $(\theta^L_q, \pi^L_q) \leftarrow \text{OPTIMIZATION}(0, \pi^{init}_q, b_{-J}(\mathfrak{p}), \mathfrak{p}, \min)$ // Solve (21)–(22) for $\min \phi$ in the sample
    end for
    if type = subsample then
        for $q = 1, \dots, \#\boldsymbol{\pi}^{init}$ do
            $(\theta^U_q, \pi^U_q) \leftarrow \text{OPTIMIZATION}(0, \pi^q_U, b_{-J} - b^{ctr}_{-J}(\theta^U_q), \mathfrak{p}, \max)$ // Solve (21)–(22) for $\max \phi$ in the subsample, with recentering
            $(\theta^L_q, \pi^L_q) \leftarrow \text{OPTIMIZATION}(0, \pi^q_L, b_{-J} - b^{ctr}_{-J}(\theta^L_q), \mathfrak{p}, \min)$ // Solve (21)–(22) for $\min \phi$ in the subsample, with recentering
        end for
    end if
    $\theta^U(0) \leftarrow \max_q \theta^U_q$;   $\pi^U(0) \leftarrow \arg\max_q \theta^U_q$;   $\theta^L(0) \leftarrow \min_q \theta^L_q$;   $\pi^L(0) \leftarrow \arg\min_q \theta^L_q$ // Select best outcomes from different initial values
    if type = sample then
        for $k = 1, \dots, \#\boldsymbol{\epsilon}$ do    // Solve optimizations in the grid set $\boldsymbol{\epsilon}$
            $(\theta^U(\epsilon_k), \pi^U(\epsilon_k)) \leftarrow \text{OPTIMIZATION}(\epsilon_k, \pi^U(\epsilon_{k-1}), b_{-J}), \mathfrak{p}, \max)$
            $(\theta^L(\epsilon_k), \pi^L(\epsilon_k)) \leftarrow \text{OPTIMIZATION}(\epsilon_k, \pi^L(\epsilon_{k-1}), b_{-J}), \mathfrak{p}, \min)$
        end for
    else type = subsample
        for $k = 1, \dots, \#\boldsymbol{\epsilon}$ do    // Solve optimizations in the grid set $\boldsymbol{\epsilon}$
            $(\theta^U(\epsilon_k), \pi^U(\epsilon_k)) \leftarrow \text{OPTIMIZATION}(\epsilon_k, \pi^U(\epsilon_{k-1}), b_{-J} - b^{ctr}_{-J}(\theta^U(\epsilon_{k-1})), \mathfrak{p}, \max)$
            $(\theta^L(\epsilon_k), \pi^L(\epsilon_k)) \leftarrow \text{OPTIMIZATION}(\epsilon_k, \pi^L(\epsilon_{k-1}), b_{-J} - b^{ctr}_{-J}(\theta^L(\epsilon_{k-1})), \mathfrak{p}, \min)$
        end for
    end if
    return $(\theta^U(\cdot)$, $\pi^U(\cdot)$, $\theta^L(\cdot)$, $\pi^L(\cdot))$
end function
```

Algorithm 3 Auxiliary Functions

```
function OPTIMIZATION($\epsilon$, $\pi^{init}$, $b_{-J}$, $\mathfrak{p}$, type)        // Compute the max and the min of $\theta$
    if type = max then
        $\theta \leftarrow \max_{\pi}\ \phi(\tilde{p}, \pi; \mathfrak{p})$ subject to (31) with $\pi^{init}$ as starting value for the optimization routine.
        $\pi \leftarrow$ argmax of previous optimization
    else if type = min then
        $\theta \leftarrow \min_{\pi}\ \phi(\tilde{p}, \pi; \mathfrak{p})$ subject to (31) with $\pi^{init}$ as starting value for the optimization routine.
        $\pi \leftarrow \arg\min$ of previous optimization
    end if
    return $(\theta, \pi)$
end function
```

```
function $\mathcal{J}(\theta^U(\cdot)$, $\theta^L(\cdot)$, $\boldsymbol{\epsilon}$, type)        // Compute the piecewise constant functions $J\colon \Theta \to \mathbb{R}$
    if $\theta^U(\epsilon_{k-1}) < \theta \leq \theta^U(\epsilon_k)$ or $\theta^L(\epsilon_{k-1}) > \theta \geq \theta^L(\epsilon_k)$ for some $k$ then
        if type = above then
            $J(\cdot) \leftarrow \epsilon_k$
        else if type = below then
            $J(\cdot) \leftarrow \epsilon_{k-1}$
        end if
    else if $\theta^L(0) \leq \theta \leq \theta^U(0)$ then
        $J(\cdot) \leftarrow 0$
    end if
    return $J(\cdot)$
end function
```

```
function RECENTERING($\theta^U(\cdot)$, $\pi^U(\cdot)$, $\theta^L(\cdot)$, $\pi^L(\cdot)$, $\boldsymbol{\epsilon}$, $\mathfrak{p}$)        // Compute the piecewise constant
function $b^{ctr}_{-J}\colon \Theta \to \mathbb{R}^{X(A-1)}$
    $b_{-J} \leftarrow -\mathbf{M}(\mathfrak{p})\boldsymbol{\psi}(\mathfrak{p})$        // Compute vector $b_{-J}$, based on equation (4)
    if $\theta^U(\epsilon_{k-1}) < \theta \leq \theta^U(\epsilon_k)$ for some $k$ then
        $b^{ctr}_{-J}(\theta) \leftarrow b_{-J} - \mathbf{M}\pi^U(\epsilon_k)$
    else if $\theta^L(0) \leq \theta \leq \theta^U(0)$ then
        $b^{ctr}_{-J}(\theta) \leftarrow 0$
    else if $\theta^L(\epsilon_{k-1}) > \theta \geq \theta^L(\epsilon_k)$ for some $k$ then
        $b^{ctr}_{-J}(\theta) \leftarrow b_{-J} - \mathbf{M}\pi^L(\epsilon_k)$
    end if
    return $b^{ctr}_{-J}(\cdot)$
end function
```

C.2 Choice of Tuning Parameters

Our inference algorithm requires setting several tuning parameters, which we now discuss.

Initial Values for Optimization. Here, we focus on the choice of the initial values for the optimization problems (30)–(31), presented in the main text. We advocate the standard practice of using several different starting values to mitigate the risks associated with local optima. We suggest starting with several randomly generated points $v_l \in \mathbb{R}^{(A+1)X}$ and find their projections onto the identified set by solving, for each point,[36]

$$\min_{\substack{\pi\in\mathbb{R}^{(A+1)X}:\\ R^{eq}\pi=r^{eq}, R^{iq}\pi\leq r^{iq},\\ \widehat{\mathbf{M}}_N\pi=b_{-J}(\widehat{\mathfrak{p}}_N)}} [v_l-\pi]'[v_l-\pi].$$

We implement the same procedure in the main sample and in each subsample. Importantly, this set of initial values is used only to solve for the identified set – i.e., for the problem with $\epsilon_0 = 0$. For any relaxed problem with $\epsilon_k > 0$, we recommend employing the solution to the optimization problem with ϵ_{k-1} as the initial value. The continuity structure of our problem indicates that the solution to the relaxed problem ϵ_{k-1} is an excellent initial point for the subsequent relaxed optimization ϵ_k, for all k.

Grid Set. When it comes to choosing the grid points, $0 \equiv \epsilon_0 < \epsilon_1 < ... < \epsilon_K \equiv \epsilon_{max}$, we suggest trying different grids in the main sample to determine the largest value of ϵ and the number of points K before subsampling. Specifically, the largest value for the grid, ϵ_{max}, should generate an (outer) set for θ, $[\theta^L(\epsilon_{max}), \theta^U(\epsilon_{max})]$, that is large enough to almost certainly include the confidence set CS. To achieve this, we recommend computing the identified set $[\theta^L(0), \theta^U(0)]$ first and then increasing ϵ until the set $[\theta^L(\epsilon_{max}), \theta^U(\epsilon_{max})]$ is either uninformative or sufficiently large based on the sample size. For example, suppose θ is the average probability of entry in a dynamic entry/exit model – so that $\theta \in \Theta = [0, 1]$ – and the estimated identified set is $\widehat{\mathbf{\Theta}}^I = [0.2, 0.3]$. To be conservative, we may choose an ϵ_{max} such that $[0.01, 0.9] \subseteq [\theta^L(\epsilon_{max}), \theta^U(\epsilon_{max})] \subseteq [0, 1]$. With a large sample size, we may choose a smaller ϵ_{max}, e.g., a value such that $[0.1, 0.4] \subseteq [\theta^L(\epsilon_{max}), \theta^U(\epsilon_{max})]$ – i.e., an outer set that is much closer to the estimated identified set. The main trade-off when choosing ϵ_{max} is between greater computational cost (from a larger ϵ_{max}) and not covering the CS with the initial outer set (from a too-small ϵ_{max}). Indeed, if ϵ_{max} is too small, the inference procedure generates a confidence set that is the intersection between the correct confidence set and $[\theta^L(\epsilon_{max}), \theta^U(\epsilon_{max})]$. If the researcher finds that one of the confidence set's extremes is equal to $\theta^L(\epsilon_{max})$ or $\theta^U(\epsilon_{max})$, it is clear that ϵ_{max} should be increased and the inference procedure

[36] This procedure assumes the identified set is non-empty. This can be checked by solving the quadratic problem

$$\min_{\substack{\pi\in\mathbb{R}^{(A+1)X}:\\ R^{eq}\pi=r^{eq}, R^{iq}\pi\leq r^{iq}}} [b_{-J}(\widehat{\mathfrak{p}}_N) - \widehat{\mathbf{M}}_N\,\pi]'\,\widehat{\Omega}_N\,[b_{-J}(\widehat{\mathfrak{p}}_N) - \widehat{\mathbf{M}}_N\,\pi].$$

If a solution π to this problem satisfies $b_{-J}(\widehat{\mathfrak{p}}_N) = \widehat{\mathbf{M}}_N\,\pi$, then we know that the identified set is non-empty. Otherwise, the identified set is empty.

performed for the new points in the grid.

Regarding the number of points K, although we acknowledge that it is not trivial to provide precise guidance since the proper choice depends on how simple/complex the particular computational problem is and how quickly the function $J(\theta)$ increases as θ "moves away" from the identified set, we note that, in our experience, starting with about 50 to 100 points for the grid generates almost the same results as finer grids. Noticing that our procedure, based on approximations to $\widehat{J}_N$ and $\widehat{J}^*_{h_N}$ from below and from above, respectively, is conservative for coarse grids in finite samples, a too coarse grid generates a confidence set that is larger than necessary. Consequently, increasing the number of grid points K can only reduce the length of the confidence set, while preserving the (minimum) nominal size. Hence, one can begin with a relatively coarse grid and gradually increase the number of grid points until the confidence set stabilizes.

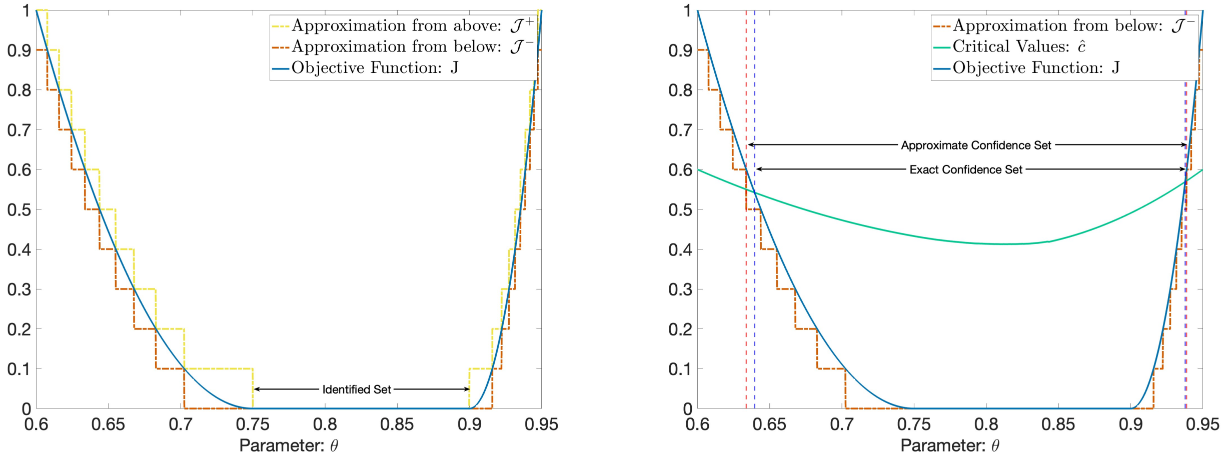
Finally, once the researcher has determined ϵ_{max} and K, we recommend using an equidistant grid with K points from 0 to ϵ_{max} for the main sample, and a grid with K points from 0 to $\frac{N}{h_N} \times \epsilon_{max}$ for the subsample. That is because failing to rescale the grid in the subsamples may lead to the lack of overlapping between the values that the test statistic $N\widehat{J}$ may take in the sample and the distribution of the test statistic in the subsamples $h_N\widehat{J}^*$, since $N \gg h_N$, which would invalidate the approximation to the critical value $\widehat{c}_{1-\alpha}$.

Approximations $\mathscr{J}^-$ and $\mathscr{J}^+$. As noted earlier, we recommend approximating $\widehat{J}_N$ from below and $\widehat{J}^*_{h_N}$ from above to be conservative in finite samples when the grid set chosen is coarse. Here, we illustrate this rationale using Figure C1, constructed based on the model presented in the Monte Carlo section, using the set of Restrictions 1; see Appendix D. Specifically, Figure C1, panel (a), shows the step-function approximations on a grid set with $K = 10$ equally-distant points and $\epsilon_{max} = 1$. Panel (b) presents the (conservative) approximated confidence set based on the approximation to $\widehat{J}_N$ from below. (As an aside, it is worth noting that the true critical value function in the figure, $\widehat{c}_{1-\alpha}(\theta)$, varies with θ, pointing to the importance of treating this function carefully in the inference procedure.)

C.3 Computational Costs

We now turn to the computational costs of our inference procedure. As mentioned previously, our algorithm is based on solving problem (30)–(31) multiple times. This accounts for the majority of the computational costs. In the absence of the equality restrictions (8), the computational cost of solving this problem grows roughly as X^2, which is consistent with our experience and Monte Carlo simulations.

Incorporating the equality restrictions (8) reduces the computational cost by decreasing the dimension of the search. Our Monte Carlo simulations are consistent with this: as we add Restriction 3 (i.e., scrap values do not depend on the exogenous states W), the average time required to perform inference drops substantially for any sample size when compared to the average time when only Restrictions 1 and 2 are imposed.



(a) J function and approximations, $\mathcal{J}^-$ and $\mathcal{J}^+$

(b) Approximated and Exact Confidence Sets

Figure C1: Approximating Functions and Confidence Sets.

Shape restrictions (9), however, do not necessarily reduce computational costs. In fact, in our experience, they can increase the likelihood of the algorithm stopping at a local optimum. Thus, we recommend trying multiple starting points, as discussed previously, especially when many inequality restrictions are used. Once again, this is consistent with our Monte Carlo results: adding Restrictions 2 (which only includes shape restrictions) on top of Restrictions 1 can reduce the average time for estimation and inference, though not always.

D A Monte Carlo Study

In this section, we present a Monte Carlo study to illustrate the finite-sample performance of our inference procedure. We start with the setup, and then we show the results.

D.1 Setup

We extend the firm entry/exit problem presented in Section 5 of the main text, allowing now for a larger state space. Specifically, we assume the presence of three exogenous states, $w_t = (w_{1t}, w_{2t}, w_{3t})$, reflecting demand and supply shocks. The exogenous states are independent to each other, and each follows a discrete-AR(1) process with W support points (obtained by discretizing latent normally-distributed AR(1) processes). The (residual) inverse demand function is linear, $P_t = \bar{w} + w_{1t} + w_{2t} - \eta Q_t$, where P_t is the price of the product, Q_t is the quantity demanded, $\bar{w}$ is the intercept, w_{1t} and w_{2t} are demand shocks, and η is the slope. We assume constant marginal costs mc_t (i.e., mc_t does not depend on Q_t), and let the supply shocks w_{3t} affect marginal costs. To simplify, we just take $mc_t = w_{3t}$. Variable profits are then $vp_t = (\bar{w} + w_{1t} + w_{2t} - mc_t)^2/4\eta$. The idiosyncratic shocks ε follow the type 1 extreme value distribution. The model parameters are presented in Table D1.

Table D1: Parameters of the Monte Carlo Data Generating Process

Demand Function:	$\bar{w}$	6.8	$w_{1t} \sim$ Normal AR(1):	ρ_{01}	0
	η	4		ρ_{11}	0.75
				σ_1^2	0.02
Payoff Parameters:	s	4.5	$w_{2t} \sim$ Normal AR(1):	ρ_{02}	0
	ec	5		ρ_{12}	0.75
	fc	0.5		σ_2^2	0.025
Scale parameter:	σ	1	$w_{3t} \sim$ Normal AR(1):	ρ_{03}	0
				ρ_{13}	0.75
Discount Factor:	β	0.9		σ_3^2	0.03

The counterfactual we consider is the same as in Section 5: a subsidy that reduces entry costs by 20%. The target parameter θ is the long-run average probability of staying in the market given the subsidy, where the long-run average is based on the ergodic distribution of the state variables; the specific formula for θ is provided in Appendix F (but note that here we do not take the difference between the counterfactual and the baseline average probabilities).

In order to analyze the sensitivity of the target parameter θ to alternative model restrictions, we follow the example again and impose Restrictions 1–3, as explained in Section 5.[37]

We generate 1000 Monte Carlo replications for each of the following sample sizes: the small sample, with $N = 100$ firms on separated (independent) markets and $T = 5$ time periods, and the large sample, with $N = 1000$ firms and $T = 15$ time periods. For the first sample period, the value of the state variables are drawn from their steady-state distributions. Given that each exogenous state variable w_{jt} can take W values, the dimension of the state space is $X = 2 \times W^3$. We consider three sizes for the state space: $X = 16$, 54 and 250, which correspond to $W = 2$, 3 and 5. The choices of the state space were dictated by the sample size, not by computational constraints, given that the method makes use of a frequency, or a nonparametric estimator for the CCP in the first stage. As discussed in the main text and in Section B of this appendix, it is feasible to solve the optimization problem (21)–(22) for state spaces that are larger in size.

In each sample, we estimate the lower and upper bounds for the target parameter, θ^L and θ^U, by solving the minimization and maximization problems (21)–(22). We estimate CCPs using frequency estimators, and we use the true transition matrix F, both in calculating test statistics and critical values. (The results do not change significantly when we estimate transition probabilities as well.) We solve the problem (21)–(22) using the Knitro MATLAB function. We randomly generated a set of 25 initial values

[37] When we impose Restriction 3, we replace the inequalities defined in Restriction 1 by their average versions. This does not affect the identified set, but it improves the finite-sample behavior of the estimators when the sample size is small and the state space is large.

for the optimizations as discussed in Appendix C. We provide the analytical gradient of ϕ as explained in Appendix H.

We specify $\widehat{\Omega}_N$ to be a diagonal matrix with diagonal terms given by the square-root of the ergodic distribution of the exogenous state variables, implied by the transition process F^w. We opt for this weighting matrix so that deviations on more visited states receive greater weights and are, therefore, considered more relevant.

We calculate 90% confidence sets for θ using the procedure described in Section 6 of the main text and in Section C of this appendix. Specifically, in each sample, we approximate $\widehat{J}_N(\theta)$ from below using a grid with a range from $\epsilon_0 = 0$ to $\epsilon_{max} = 0.1$ in an equally spaced grid with $K = 50$ points. For each sample, we generate 1000 subsamples with size that is approximately $h_N \approx N^{\frac{2}{3}}$. More precisely, we implement a standard i.i.d. subsampling, resampling firms over the full time period: For the small sample we draw 22 firms randomly, and for the large sample, we draw 100 firms. For the subsamples, we used a grid with a range from $\epsilon_{min} = 0$ to $\epsilon_{max} = 1$ in an equally spaced grid with $K = 50$ points, and we have approximated $\widehat{J}^*_{h_N}$ from above. The computations were run on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University.

D.2 Monte Carlo Results

Main Results. We now discuss the results of the Monte Carlo simulations. In the baseline scenario, the long-run average probability that the firm stays in the market is 90.5%, while the long-run average probability of being active reduces to 83.3% in the counterfactual scenario (so that $\theta = 0.833$). The impact of the entry subsidy is to reduce the long-run average by 7.2 percentage points. Similar to the example presented in the main paper, the entry subsidy increases the exit rate of forward-looking firms, which translates into firms staying less often in the market in the steady state. This result is invariant to the alternative discretizations of the state space, since the discretizations are performed on the same underlying AR(1) processes.

Table D2 presents the Monte Carlo results. The top, middle, and bottom panels show the results for the alternative state spaces: small ($X = 16$), medium-sized ($X = 54$), and large ($X = 250$), respectively. In each panel, the top subpanel presents the results for the small sample ($N = 100, T = 5$), and the bottom subpanel, for the large sample ($N = 1000, T = 15$). In each subpanel, we show for each alternative set of Restrictions 1–3, (i) the populational (true) identified set, (ii) the average estimates of the lower and upper bounds, θ^L and θ^U, (iii) the average bias of the estimated bounds, (iv) the average endpoints and the average length of the 90% confidence sets, (v) the coverage probability of the confidence sets, and (vi) the average time taken to estimate θ^L and θ^U (in seconds), and the average time taken to compute the confidence intervals (in minutes).

The identified sets under the alternative Restrictions 1–3 are all compact intervals containing the true θ (Proposition 5), and vary slightly with the size of the state space. To be concrete, for the medium-sized state space ($X = 54$), Restriction 1 alone is highly informative: the counterfactual long-run average probability of being active is between 75% and 90.5%. It does however include the baseline probability (at the upper end of the interval). Adding Restriction 2 reduces the upper bound to 87.8%, which suffices to identify the sign of the impact of the subsidy. Adding Restriction 3 pushes the upper bound further down to 86.8%.

In all cases, the estimated lower and upper bounds of the identified sets appear to be consistent, with smaller biases in larger samples. The coverage probabilities of the confidence sets converge to the nominal level 90%, as expected (Theorem 1). And the confidence sets' average lengths are wider (though not substantially) than the length of the true identified sets, for all sample sizes and state spaces. E.g., in the small state space case and large sample, the average length of the confidence set is 0.1774 under Restriction 1, while the length of the (true) identified set is just 0.1533; and in the large state space and small sample, the average length of the confidence set under the same restriction is 0.2512.

Naturally, the finite sample performance of our inference procedure depends on both the state space and the sample size. In the larger state space cases, we obtain slightly greater average biases for the point estimates. These are expected: larger state spaces imply less (effective) degrees of freedom, as the number of model parameters increases with the state space (recall that π is an $(A+1)X$ vector).

In terms of the computer time required to solve the minimization and maximization problems (21)–(22), it takes approximately 0.04 seconds to solve both optimization problems under Restrictions 1 and 1–2, and that time is reduced to just 0.01 seconds under Restrictions 1–3, in the small state space case. Subsampling is computationally intensive but feasible: for the same state space, the average time required to run it varies from two minutes under Restriction 1 to just one minute under Restrictions 1–3.

As expected, it takes longer to solve (21)–(22) when the state space is larger. E.g., under Restriction 1, it takes approximately 0.3 seconds on average in the medium-sized state space case ($X = 54$), and approximately 6 seconds on average in the large state space case ($X = 250$). It also takes longer to run the subsampling procedure: between 6 and 26 minutes on average in the medium-sized state space, and between 125 and 521 minutes on average in the large state space, depending on the sample size and the restrictions imposed. It is important to stress, however, that the average computer time here is based on a sequential implementation of subsampling, which does not take advantage of parallelization.

Robustness Analysis. Next, we investigate the robustness of our results to (a) the choice of the subsample size, (b) the selection of the grid points, and (c) the approximations to our test statistics, $J_N(\theta)$ (whether from below or from above). To facilitate the comparisons, we focus only on the case with medium-sized state space ($X = 54$) and large sample size ($N = 1000, T = 15$).

Table D3 presents the results when we select the subsample size to be $h_N = N/4$, following the suggestion of Ciliberto and Tamer (2009). This implies that we randomly sample 250 firms instead of sampling $h_N = N^{\frac{2}{3}} = 100$ firms, as suggested by Bugni (2016) and presented in Table D2. All results, including the coverage probability and computational time, are strikingly similar to each other.

Table D4 presents the Monte Carlo results obtained by varying the grid points and focusing on Restriction 1 for simplicity. We consider a range from $\epsilon_0 = 0$ to $\epsilon_{max} = 0.1$ using an equally spaced grid and vary the number of points K. Specifically, we study four scenarios: (i) a fine grid with $K = 100$ points, (ii) our main grid set with $K = 50$ points, which was previously presented in Table D2 and is replicated here for convenience, (iii) a coarser grid with $K = 25$ points, and (iv) an even coarser grid with $K = 12$ points.

The results confirm the intuition that finer grids provide more accurate outcomes at the cost of increased computational time. Notably, the finer grids ($K = 100$ and $K = 50$) yield correct coverage probabilities, while the coarser grids ($K = 25$ and $K = 12$) result in more conservative confidence intervals with coverage probabilities above 90%. As for computational time, using finer grids takes longer to compute the confidence sets compared to coarser grids. For instance, the average computation time is 32 minutes for the finest grid ($K = 100$) and only 10 minutes for the coarsest grid ($K = 12$). However, it is important to note that the computational time increases less than linearly with the number of grid points K. This is likely due to the similarity between the relaxed problems based on ϵ_k and ϵ_{k+1}, $k = 0, \ldots, K-1$, where the solution for ϵ_k provides an excellent initial value for the problem with ϵ_{k+1}. Consequently, solving a larger number of relaxed problems as the grid gets finer does not significantly add to the computational cost. (Note that, in contrast, when ϵ_k and ϵ_{k+1} are further apart, in a coarser grid, then ϵ_k is not as good an initial value for the relaxed problem with ϵ_{k+1} as when ϵ_k and ϵ_{k+1} are closer together.) Given these findings, we recommend using finer grid points in practice due to their ability to provide more accurate outcomes, despite the moderately longer computational time.

Table D5 presents the results when we explore different ways of approximating our test statistics, focusing as before on Restiction 1. Recall that, in the main sample, we approximate $\widehat{J}_N$ from below using $\mathcal{J}^-$, while in the subsamples we approximate $\widehat{J}^*_{h_N}$ from above using $\mathcal{J}^+$. This approach is adopted to avoid undercoverage in finite samples when the set of grid points is coarse. In Table D5, we now showcase the results when we approximate our test statistics from above in *both* the main sample (leading to undercoverage) and all subsamples (leading to overcoverage). Our goal is to investigate which approximation dominates. Notably, our original grid ($K = 50$) and the finer grid ($K = 100$) yield coverage probabilities that align with the correct nominal size. However, the use of coarser grids ($K = 12$ or $K = 25$) results in undercoverage. For $K = 12$, in particular, the undercoverage can be substantial (0.845), while our proposed solution is reasonably conservative in this case (0.942, presented in Table D4). This undercoverage appears to happen because $N \gg h_N$ implies that the approximation error to the test statistic in the main sample ($N \times \widehat{J}_N$) dominates the approximation error to the critical value (based on $h_N \times \widehat{J}^*_{h_N}$).

Table D2: Monte Carlo Results

Target Parameter: θ = Long-run Average Probability of Being Active			
Small State Space: $X = 16$			
$T = 5, N = 100$	Restrictions 1	Restrictions 1–2	Restrictions 1–3
True Identified Set	[0.7500, 0.9033]	[0.7500, 0.8765]	[0.7500, 0.8661]
Average Estimated Bounds	[0.7582, 0.9036]	[0.7580, 0.8722]	[0.7580, 0.8655]
Average Bias	[0.0082, 0.0003]	[0.0080, -0.0038]	[0.0080, -0.0010]
Confidence Sets: Average Endpoints	[0.6761, 0.9205]	[0.6760, 0.8943]	[0.6757, 0.8854]
Confidence Sets: Average Length	0.2444	0.2183	0.2097
Coverage Probability (90% nominal)	0.8930	0.8870	0.8700
Time Estimation (sec)	0.04	0.04	0.01
Time Inference (min)	2	2	1
$T = 15, N = 1000$	Restrictions 1	Restrictions 1–2	Restrictions 1–3
True Identified Set	[0.7500, 0.9033]	[0.7500, 0.8765]	[0.7500, 0.8661]
Average Estimated Bounds	[0.7507, 0.9034]	[0.7507, 0.8762]	[0.7507, 0.8657]
Average Bias	[0.0007, 0.0001]	[0.0007, -0.0003]	[0.0007, -0.0004]
Confidence Sets: Average Endpoints	[0.7302, 0.9076]	[0.7302, 0.8806]	[0.7302, 0.8712]
Confidence Sets: Average Length	0.1774	0.1504	0.1410
Coverage Probability (90% nominal)	0.8900	0.8880	0.9060
Time Estimation (sec)	0.04	0.04	0.01
Time Inference (min)	2	2	1
Medium-sized State Space: $X = 54$			
$T = 5, N = 100$	Restrictions 1	Restrictions 1–2	Restrictions 1–3
True Identified Set	[0.7501, 0.9052]	[0.7501, 0.8781]	[0.7501, 0.8684]
Average Estimated Bounds	[0.7595, 0.9034]	[0.7591, 0.8709]	[0.7591, 0.8646]
Average Bias	[0.0094, -0.0018]	[0.0090, -0.0072]	[0.0090, -0.0038]
Confidence Sets: Average Endpoints	[0.6594, 0.9248]	[0.6587, 0.9066]	[0.6587, 0.8939]
Confidence Sets: Average Length	0.2654	0.2479	0.2352
Coverage Probability (90% nominal)	0.9230	0.918	0.8970
Time Estimation (sec)	0.32	0.36	0.03
Time Inference (min)	26	20	11
$T = 15, N = 1000$	Restrictions 1	Restrictions 1–2	Restrictions 1–3
True Identified Set	[0.7501, 0.9052]	[0.7501, 0.8781]	[0.7501, 0.8684]
Average Estimated Bounds	[0.7506, 0.9052]	[0.7506, 0.8776]	[0.7506, 0.8680]
Average Bias	[0.0005, 0.0000]	[0.0005, -0.0005]	[0.0005, -0.0004]
Confidence Sets: Average Endpoints	[0.7279, 0.9099]	[0.7279, 0.8829]	[0.7279, 0.8751]
Confidence Sets: Average Length	0.1820	0.1550	0.1472
Coverage Probability (90% nominal)	0.9080	0.9190	0.9110
Time Estimation (sec)	0.28	0.22	0.03
Time Inference (min)	22	16	6
Large State Space: $X = 250$			
$T = 5, N = 100$	Restrictions 1	Restrictions 1–2	Restrictions 1–3
True Identified Set	[0.7506, 0.9054]	[0.7506, 0.8785]	[0.7506, 0.8688]
Average Estimated Bounds	[0.7624, 0.9024]	[0.7615, 0.8746]	[0.7603, 0.8648]
Average Bias	[0.0118, -0.0030]	[0.0109, -0.0039]	[0.0097, -0.0040]
Confidence Sets: Average Endpoints	[0.6741, 0.9253]	[0.6730, 0.9129]	[0.6711, 0.8980]
Confidence Sets: Average Length	0.2512	0.2399	0.2269
Coverage Probability (90% nominal)	0.8780	0.8860	0.8900
Time Estimation (sec)	6	6	1
Time Inference (min)	521	433	212
$T = 15, N = 1000$	Restrictions 1	Restrictions 1–2	Restrictions 1–3
True Identified Set	[0.7506, 0.9054]	[0.7506, 0.8785]	[0.7503, 0.8688]
Average Estimated Bounds	[0.7524, 0.9058]	[0.7524, 0.8787]	[0.7524, 0.8690]
Average Bias	[0.0018, 0.0004]	[0.0018, 0.0002]	[0.0018, 0.0002]
Confidence Sets: Average Endpoints	[0.7316, 0.9099]	[0.7316, 0.8848]	[0.7321, 0.8761]
Confidence Sets: Average Length	0.1783	0.1532	0.1440
Coverage Probability (90% nominal)	0.898	0.8830	0.9040
Time Estimation (sec)	6	5	1
Time Inference (min)	446	399	125

Note: T = number of periods, N = number of markets, X = number of states.

Table D3: Monte Carlo Results – Alternative subsample size, $h_N = \frac{N}{4}$

Medium-sized State Space: $X = 54$			
$T = 15, N = 1000$	Restrictions 1	Restrictions 1–2	Restrictions 1–3
True Identified Set	[0.7501, 0.9052]	[0.7501, 0.8781]	[0.7501, 0.8684]
Average Estimated Bounds	[0.7506, 0.9052]	[0.7506, 0.8776]	[0.7506, 0.8680]
Average Bias	[0.0005, 0.0000]	[0.0005, -0.0005]	[0.0005, -0.0004]
Confidence Sets: Average Endpoints	[0.7284, 0.9097]	[0.7284, 0.8820]	[0.7284, 0.8748]
Confidence Sets: Average Length	0.1813	0.1536	0.1464
Coverage Probability (90% nominal)	0.9020	0.8940	0.9060
Time Estimation (sec)	0.27	0.23	0.03
Time Inference (min)	20	15	6

Note: T = number of periods, N = number of markets, X = number of states.

Table D4: Monte Carlo Results – Changing the number of gridpoints

Medium-sized State Space: $X = 54$				
$T = 15, N = 1000$	Restriction 1	Restriction 1	Restriction 1	Restriction 1
Gridpoints K	100	50	25	12
True Identified Set	[0.7501, 0.9052]	[0.7501, 0.9052]	[0.7501, 0.9052]	[0.7501, 0.9052]
Confidence Sets: Average Endpoints	[0.7280, 0.9097]	[0.7279, 0.9099]	[0.7269, 0.9101]	[0.7255, 0.9105]
Confidence Sets: Average Length	0.1817	0.1820	0.1832	0.1850
Coverage Probability (90% nominal)	0.9060	0.9080	0.9200	0.9420
Time Inference (min)	32	22	15	10

Note: T = number of periods, N = number of markets, X = number of states.
Our results use a gridchoice with $K = 50$.

Table D5: Monte Carlo Results – Change number of gridpoints & approximate $\widehat{J}$ from above.

Medium-sized State Space: $X = 54$				
$T = 15, N = 1000$	Restriction 1	Restriction 1	Restriction 1	Restriction 1
Gridpoints K	100	50	25	12
True Identified Set	[0.7501, 0.9052]	[0.7501, 0.9052]	[0.7501, 0.9052]	[0.7501, 0.9052]
Confidence Sets: Average Endpoints	[0.7284, 0.9098]	[0.7286, 0.9097]	[0.7292, 0.9095]	[0.7313, 0.9091]
Confidence Sets: Average Length	0.1814	0.1811	0.1803	0.1778
Coverage Probability (90% nominal)	0.9010	0.8970	0.8880	0.8450
Time Inference (min)	32	22	15	10

Note: T = number of periods, N = number of markets, X = number of states.
Our results use a gridchoice with $K = 50$.