Moment Inequalities for Entry Games with Heterogeneous Types

Christian Bontemps * Rohit Kumar † Max Lesellier ‡

November 14, 2022
link to the last version

Abstract

Following Bresnahan and Reiss (1991a), Bresnahan and Reiss (1991b) and Berry (1992), entry games have become a popular model in the empirical industrial organization literature. They enable researchers to study different features of an industry with easy-to-obtain data on entry. In this paper, we provide new tools to simplify the estimation of entry games when the equilibrium selection mechanism is unrestricted. In particular, we develop an algorithm that allows us to recursively select a relevant subset of inequalities and compute the theoretical upper bounds on the probability of each outcome (without having to resort to simulation-based methods). We also propose a new testing procedure that is asymptotically pivotal by smoothing the set defined by the moment inequalities. We show that this new estimation procedure can seamlessly accommodate covariates, including continuous ones. We conduct full-scale Monte Carlo simulations to assess the performance of our new estimation procedure.

Keywords: empirical entry games, moment inequalities, core determining class, smoothing.

*ENAC and Toulouse School of Economics. Email: christian.bontemps@tse-fr.eu
†Indian Institute of Technology Delhi. Email: rohitk@iitd.ac.in
‡Toulouse School of Economics. Email: max.lesellier@ut-capitole.fr

1 Introduction

In the wake of the seminal contributions by Bresnahan and Reiss (1991a), Bresnahan and Reiss (1991b) and Berry (1992), entry games have become a popular model in the empirical industrial organization literature (IO). They allow researchers to study the determinants of firm profitability and the degree of competition from data on entry, which is usually easy to collect. Entry games can also serve as a building block to multi-stage games involving for instance price competition (eg: Eizenberg (2014), Ciliberto et al. (2021)...). Among the most influential applications, one can mention the analysis of competition, market structure and regulation in various industries including airlines (Ciliberto and Tamer (2009), Berry (1992)), retailers (Cleeren et al. (2010), Aradillas-Lopez and Rosen (2022), Andrews et al. (2004), Grieco (2014)...), motels (Mazzeo (2002)) and fast food restaurants (Toivanen and Waterson (2005)).

From a methodological perspective, this paper provides a novel estimation strategy for static entry games of complete information, which significantly simplifies the estimation procedure when no restrictions are imposed on the equilibrium selection mechanism. It is a well known difficulty that in the absence of a known equilibrium selection mechanism, the estimation is substantially complicated by the presence of multiple equilibria. Namely, there are regions in the space of unobservable shocks in which the entry game predicts multiple equilibria and yet the econometrician only observes one outcome. Without additional information on the equilibrium selection mechanism, the model is said to be incomplete and the econometrician cannot rely on standard estimation techniques. To tackle this problem, the literature proposes various solutions, which can be divided in two distinct categories. The first approach consists in restricting the equilibrium selection mechanism so as to complete the model. There are various ways of proceeding. The econometrician can impose an order of entry (Mazzeo (2002), Cleeren et al. (2010)...). Bjorn and Vuong (1984) suggests to randomly draw an equilibrium out of the multiple potential equilibria. Grieco (2014) and Bajari et al. (2010) explicitly model the equilibrium selection mechanism. Building on the vast literature related to set identification initiated by Manski (1995), the second generic solution consists in characterizing the set of parameters which can generate the observed data without restricting the equilibrium selection mechanism (among prominent examples of this approach in the context of games, see

Ciliberto and Tamer (2009), Beresteanu et al. (2011), Galichon and Henry (2011), Beresteanu et al. (2012), Bontemps and Kumar (2020), Aradillas-Lopez and Rosen (2022), Chesher and Rosen (2019), magnolfironcoroni2022). Despite the risk of misspecification implied by incorrect restrictions on the equilibrium selection mechanism, heretofore, the empirical literature has largely favored the first approach due to the relative simplicity of its implementation. The second family of solutions is less restrictive but its implementation faces major theoretical and practical challenges. First, even in seemingly harmless games, characterizing the sharp identified set (the set of admissible parameters, which satisfy all the inequalities implied by the model) can be a grueling task. (i) The number of inequalities generated by the model increases exponentially with the number of players and can quickly become overwhelming[1]. (ii) Except for a few toy models, there are no closed form expressions for the theoretical bounds implied by the model and one must resort to simulation methods, which mechanically induce biases. Additionally, the estimation of the identified set also poses many challenges. (i) The exact asymptotic distribution of the test statistic under the null depends on the set of binding moments, which is unknown to the econometrician. This seriously complicates the derivation of the critical value. The methods proposed in the literature either rely on simulation methods, which are computationally intensive[2], or upper bounds, which are conservative. Moreover, the finite sample performance of these methods is known to decrease steadily with the inclusion of many moment inequalities, which mechanically inflate the critical value. (ii) The presence of exogenous covariates (and in particular, continuous ones) complicates even more the estimation as the identified set is now characterized by conditional moment inequalities that must be converted into unconditional ones. (iii) Finally, the estimation of confidence region for the structural parameters is based on a test inversion over a grid, which can quickly become very large if the dimension of the structural parameter θ increases. Therefore, the objective of this paper is to remove or mitigate most of the difficulties exposed above, and thus, facilitate and encourage the estimation of static entry games in empirical work, while remaining agnostic about the equilibrium selection mechanism.

We now briefly summarize the main methodological improvements that we initiate in the paper.

[1] For classical entry games, with N players, the total number of inequalities is 2^{2^N}

[2] sub-sampling, bootstrap or simulation of the asymptotic distribution and these methods even if they mitigate the inferential loss, still yield conservative critical values

The baseline model we study is a generic static entry game with types, which corresponds to a generalization of entry games where some players are pooled together according to their characteristics. By regrouping some of the competitors in a same category, we can substantially increase the number of potential entrants, while keeping the number of parameters to estimate low[3]. The first part of the paper addresses the challenges related to identification. Each candidate parameter induces a graph on the set of outcomes, which is such that there is a link between two outcomes if their equilibrium regions overlap. To reduce the number of inequalities that sharply characterize the identified set, we leverage this graph over the set of outcomes induced by each candidate parameter θ. The novelty in this paper is to provide a systematic way of deriving the graph and inferring the subset of relevant inequalities. As for the computation of the theoretical bounds, we show how to derive them by exploiting the inclusion-exclusion formula and observing that intersection regions are cubes, for which the bounds can be easily derived[4]. In the second part of the paper, we tackle the issues related to estimation. To mitigate the inferential loss due to the inability to recover the exact asymptotic distribution in the context of moment inequalities, we develop an alternative approach which consists in smoothing the identified set in order to recover a test statistic with a known and pivotal asymptotic distribution. The smoothed set that we estimate is an outer set of the sharp identification set, which we make converge to the sharp identified set by letting the amount of smoothing decrease with the sample size. The general philosophy of this approach can be linked to the common bias variance trade off which appears in most econometric problems. We provide a general guideline on how to optimally choose the smoothing parameter. Last but not least, we show that this smoothing procedure facilitates the inclusion of covariates into the model, which represents a major improvement with respect to the rest of the literature.

Related literature While this paper focuses essentially on the estimation of entry games, some of the tools we develop in this paper apply more broadly to the estimation of models characterized by moment inequalities. In this sense, this paper contributes to the rich literature on conditional and unconditional moment inequalities, which includes, among others, contributions by Chernozhukov et al. (2007), Rosen (2008), Beresteanu and Molinari (2008), Andrews and Soares (2010), Bontemps

[3]Thus, the introduction of types help reduce the size of grid that we need to explore in the estimation

[4]allowing the theoretical probabilities to be easily derived by integrating over cubes

et al. (2012), Romano et al. (2014), Chernozhukov et al. (2018b), Andrews and Shi (2013), Armstrong and Chan (2016), Armstrong (2014), Molchanov and Molinari (2014), Bugni et al. (2017),Cox and Shi (2022), Chen et al. (2018), Kitamura and Stoye (2018), Kaido et al. (2019), Andrews et al. (ming), Cho and Russell (2018), Gafarov (2019), Berry and Compiani (2022).

Structure of the paper. The remainder of the paper is organized as follows. Section 2 describes the general set-up as well as the standard assumptions we impose on the model. In section 3, we characterize the identified set and we present a practical approach to the selection of relevant inequalities and the derivation of the theoretical bounds. In section 4, we present our novel estimation strategy which builds on smoothing the identified and we compare our approach with more conventional procedures. In section 5, we provide some Monte Carlo simulations to assess the performance of our estimation procedure in comparison to alternative strategies proposed in the literature. The proofs are given in the Appendix.

2 The model

We consider a flexible entry game model in the spirit of the models already developed in the literature (Berry (1992), Aradillas-Lopez and Rosen (2022) and Cleeren et al. (2010) among others). We pool some of the firms a cording to their types (or format). In this set-up, profit functions are heterogeneous across types and homogeneous within each type.[5] As we see later, there is a trade-off between the accuracy of the inference procedure and the flexibility of the model. Pooling the different competitors in types results in a substantial decrease in the number of parameters to estimate while not focusing on their true identity and, therefore, keeping a large number of potential entrants. For example, in the airline industry, small Low Cost Carriers only operate in a few markets and what matters in a given origin/destination market is how many LCCs are operating in this market rather than which LCC is operating. Similarly, in the retail industry, firms of the same format have the same business model and people go shopping at the local hypermarket or one of the closest hard-discounters, whatever their specific brand. Interestingly, models with

[5]We want to emphasize that traditional entry games in which all the players have different profit functions are simply a special case of this model in which all types can have at most 1 player (hence, in this specific case, each type represents a single player).

discrete outcomes like the one of Aradillas-Lopez and Rosen (2022) share a similar structure and the discussion below can be adapted to this case. Finally, we assume that the types of the firms are predetermined and not endogeneously chosen. If there is free-entry, this assumption is not restrictive and results like the ones in Mazzeo (2002) can be derived similarly. We describe now in more specific details our model before studying the equilibrium structure.

2.1 Payoff for entering firms

In a given market m, the profit of an entering firm of type t depends on the number of entering firms $N_{t,m}$ of each type, $t = 1, \ldots, T$, ($\mathcal{T}$ denotes the set of indices) gathered into one vector $Y = (N_{1,m}, N_{2,m}, \ldots, N_{T,m})$. It also depends on a vector of d market and type characteristics $X_{t,m}$ and a firm profit shock $\varepsilon t, m$, which is market and type specific, drawn from a parametric distribution $F_\eta(\cdot)$, $\eta \in \Lambda \subset \mathbb{R}^q$ and are independent from the characteristics. Formally, we have for each market m:

$\forall t \in \{1, \ldots, T\}$,

$$\Pi_{t,m} = \pi_t(X_{t,m}, N_{t,m}, \mathbf{N}_{-t,m}; \omega) + \varepsilon_{t,m},$$

in which the function π_t is parametrized by parameter $\omega \in \mathbb{R}^{q'}$ and $\mathbf{N}_{-t,m}$ denotes the vector of the number of entering firms of type $t' \neq t$ in market m. Observe that we keep the possibility to get heterogeneous reactions on a type t firm's profit with respect to the potential entry of firms of different types. In the following parameter θ denotes the $q + q'$ vector of parameters gathering ω and η. θ_0 is the unknown true value.

Now, we impose some restrictions on the profit function, coming from economic restrictions.

Assumption 1 *The profit is decreasing with respect to the number of competitors, i.e.,* $\forall t \in \mathcal{T}$, $\pi_t(X_m, N_{t,m}, \mathbf{N}_{-t,m}; \omega)$ *is strictly decreasing in* $N_{t,m}$ *and weekly decreasing in any* $N_{t',m}$, $t' \neq t$.

Assumption 1 uses the fact that more competitors are worse for economic profitability of a firm of a given type t.

Also, firms enter if their long run profit is weakly positive, otherwise receive a zero payoff.

We assume that firms have complete information [6] and thus observe all the profit shocks of their competitors when they decide to enter or not, contrary to the econometrician. Additionally, when making their decisions, they do not observe the decisions of the other firms, and thus all make simultaneous moves. We focus on pure strategy Nash Equilibria (NE hereafter), like most of the literature (Berry (1992), Ciliberto and Tamer (2009), Bontemps and Kumar (2020) or Aradillas-Lopez and Rosen (2022)). As it is well known, different equilibria concepts can be considered (mixed strategy or correlated equilibria), and the solutions proposed in this paper can be adapted to this new setting by adapting the set of moment inequalities which are derived (see, in particular, Beresteanu et al., 2011 or Magnolfi and Roncoroni, 2022)

Additionally, we assume the following conditions:

Assumption 2 *(Regularity assumptions on the profit functions)*

- $\forall t \in \mathcal{T}$, $\pi_t(X_{t,m}, 0, \mathbf{N}_{-t,m}; \omega) = +\infty$.
- $\forall t \in \mathcal{T}$, $\lim_{N_t \to +\infty} \pi_t(X_{t,m}, N_t, \mathbf{N}_{-t,m}; \omega) = -\infty$.
- *The distribution of the profit shocks ε, $F_\eta(\cdot)$, is absolutely continuous on $\mathbb{R}^T$ with full support and mean 0.*

Assumption 2 is standard. The first two ones are only a normalisation to calculate for the first one the probability of no entry for a given type given the number of entrants of the other types and to ensure the finiteness of Y for the second one. The third one ensures that for any X the probability to observe no entry is always strictly positive. Less restrictive assumptions can be made to ensure the same requirements like in Aradillas-Lopez and Rosen (2022) but ours is not very restrictive. In the remaining of the paper, we refer to the following model to illustrate our results. As a pedagogical example of our general model, we consider this simple 2-type game with linear profit functions and no covariates.

[6] Grieco (2014) and Bajari et al. (2006) provide identification and estimation strategies to tackle games of incomplete information.

Example 1 *Profit functions of firms of type 1 and 2 write as follows, omitting the subscript* m *for the ease of the exposition:*

$$\Pi_1 = \beta_1 - \delta_{1,1}N_1 - \delta_{2,1}N_2 + \varepsilon_1$$

$$\Pi_2 = \beta_2 - \delta_{1,2}N_1 - \delta_{2,2}N_2 + \varepsilon_2$$

where

- N_t *is the number of firms of type* $t = 1, 2$*, active in the markets*
- ε_t *unobserved heterogeneity for types* $t = 1, 2$*,*
- $\theta = (\beta_1, \beta_2, \delta_{1,1}, \delta_{2,1}, \delta_{1,2}, \delta_{2,2})$ *is the parameter of interest which we seek to identify.* δ*'s capture competition effects within each type and between types.*[7]

We allow a maximum of 3 potential entrants of each type $t = 1, 2$*. Unobserved shocks are normally distributed, with unit variance and uncorrelated.*

2.2 Equilibrium Structure

To lighten the notations, we also drop the market index m from now on. An outcome $y = (N_1, ..., N_T)$ is a NE if each type t number of entrants N_t is a best response to the other types' number of entrants, $\mathbf{N}_{-t}$. We recall that we assume free entry without loss of generality. Therefore, for each type t, we have:

- First, it is profitable for any type t firm which is entering to operate with such a market structure, i.e.,

$$\pi_t(X, N_t, \mathbf{N}_{-t}; \omega) + \varepsilon_t \geq 0.$$

- Second, an additional entrant of the same type would lead to negative profit for any type t firm, i.e.,

$$\pi_t(X, N_t + 1, \mathbf{N}_{-t}; \omega) + \varepsilon_t < 0.$$

As a result, we have the following necessary condition for $Y = (N_1, ..., N_T)$ to be a Pure Strategy Nash Equilibrium.

[7] In the pictures drawn across the text, the values chosen for the parameters are $\beta_1 = 3$, $\beta_2 = 2$, $\delta_{11} = \delta_{22} = 1.5$ and $\delta_{12} = \delta_{21} = 0.5$

Proposition 1 *An outcome $Y = (N_1, ..., N_T)$ is a NE of our game if and only if: $\forall t \in \mathcal{T}$,*

$$-\pi_t(X, N_t, \mathbf{N}_{-t}; \omega) \leq \varepsilon_t < -\pi_t(X, N_t + 1, \mathbf{N}_{-t}; \omega). \quad (1)$$

In the following, we denote this region $\mathcal{R}_\omega(X, Y)$ and we call it the region associated to the outcome Y.

Following Assumptions 1 and 2, the existence of a NE is straightforward because, given $\mathbf{N}_{-t}$, the thresholds defined in Equation (1) for ε_t are defining a sequence of non-overlapping intervals which cover $\mathbb{R}$ for a given X.

On the contrary, it is well-known that the regions $\mathcal{R}_\omega(X, Y)$, if they cover $\mathbb{R}^T$ might overlap. For example, Figure 1 displays some of the regions[8] $(\mathcal{R}(N_1, N_2))$ for our leading example.

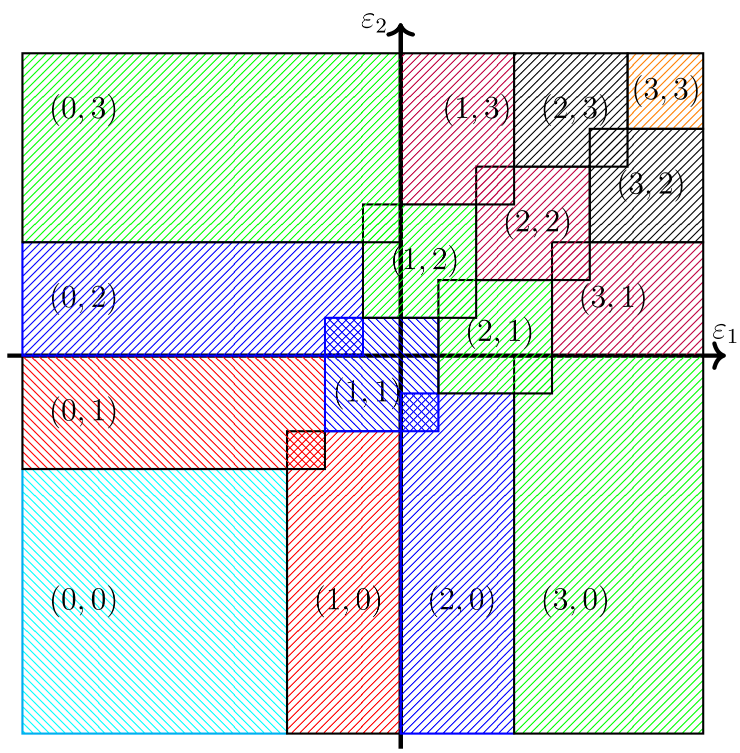


Figure 1: Equilibrium structure for $\beta_1 = 3$, $\beta_2 = 2$, $\delta_{11} = \delta_{22} = 1.5$ and $\delta_{12} = \delta_{21} = 0.5$

In Figure 1, $\mathcal{R}_\omega(1, 0)$ overlaps with $\mathcal{R}_\omega(0, 1)$, i.e., for such a draw of profit shocks, either $(1, 0)$ or $(0, 1)$ is an outcome but we can't say which one is the realized one. In other words, we have

[8]We omit X in the notation as there is no covariate in this example.

multiplicity of equilibria, i.e., there are regions of realizations of ε which do not predict a single outcome. We call them, with an abuse of langage, multiple equilibria regions. In the absence of a known equilibrium selection mechanism, there is no longer a one-to-one mapping between the set of observed outcomes and the regions of profit shocks, which prevents from using the usual identification and estimation procedures. The model is said to be *incomplete*.

The most straightforward way to circumvent the multiplicity issue is to impose restrictions on the equilibrium selection mechanism so as to ensure equilibrium uniqueness in each region of the space of unobserved heterogeneity. This is by far the approach which has gained the most traction in the empirical literature. There are various ways of restricting the equilibrium selection mechanism: Mazzeo (2002) and Cleeren et al. (2010) impose an order of entry over types, Bajari et al. (2010) explicitly models the equilibrium selection mechanism as a parametric function which can be estimated.[9] By constraining the equilibrium selection mechanism, the econometrician forces each region of the space of unobserved heterogeneity to yield a unique equilibrium. We say that imposing an equilibrium selection completes the model. The econometrician is then able to associate a well defined probability to each observed outcome.

For example, with our example, if we impose than firms of type 1 always decide first, i.e., before firms of type 2 the predictions of the model are now unique as illustrated in Figure 2. As a result, a likelihood can be derived and we can apply standard procedures. However, this strategy suffers from a huge specification risk. Another alternative, exploited in Berry (1992) in particular, is to look for combination of outcomes which are invariant in the regions of multiple equilibria. Berry (1992) shows that the number of active firms is constant at the equilibrium. Furthermore, Cleeren et al. (2010) show that, with two types and additional mild restrictions on the profit function, this remains valid. However, it is linked to particular interaction structures and it is impossible to generalize to more than two types unless imposing strong restrictions on the horizontal positioning of the different types.

This is the reason why the recent literature on moment inequalities has been used to estimate games with multiple equilibria. In fact, one can exploit the inequalities implied by the the model

[9]The set-up is slightly different to ours: mixed strategies are allowed but no types are considered and stochastic shocks are action dependent.

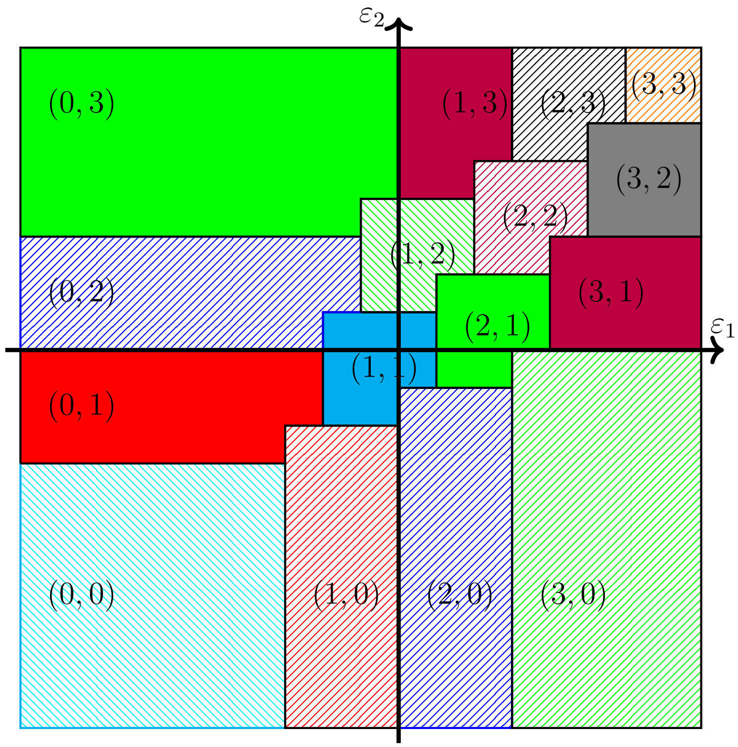


Figure 2: Equilibrium structure for $\beta_1 = 3$, $\beta_2 = 2$, $\delta_{11} = \delta_{22} = 1.5$ and $\delta_{12} = \delta_{21} = 0.5$ when firms of type 1 enter first.

while abstaining from making assumptions on the equilibrium selection mechanism. Also, this methods permits to imagine that the selection mechanism differs from one market to the other. This strategy has been extensively studied by econometricians (see Andrews et al. (2004),Ciliberto and Tamer (2009), Beresteanu et al. (2012), Galichon and Henry (2011), Bontemps and Kumar (2020)...). Andrews et al. (2004) suggests to derive an upper bound on the probability of each individual outcome. Ciliberto and Tamer (2009) improves upon Andrews et al. (2004) by computing lower bounds on the probability of each outcome. Galichon and Henry (2011) proposes a sharp characterization of the identified set which exhausts all the moments implied by the model by deriving a theoretical upper bound on each subset of the set of outcomes $\mathcal{Y}$. Beresteanu et al. (2011) generates the set of moment inequalities required to obtain a sharp characterization of the identified set but, as we show below, it becomes quickly numerically intractable. Bontemps and Kumar (2020) proposes, in a entry model à la Berry, a selection among this set of inequalities by looking for the adjacent vertex of a convex set and using only the inequalities related to this vertex. In the next section, we show how we select our inequalities by exploiting our specific structure.

3 Deriving the smallest set of moment inequalities

First, we define a few additional notations. $\mathcal{Y}$ is the set of possible outcomes of the game, ordered by the econometrician[10] and $\mathcal{X}$ the support of the exogenous covariates X. r is the cardinality of $\mathcal{Y}$. For any subset A of $\mathcal{Y}$, (i.e., $A \in \mathcal{P}(\mathcal{Y})$), $P_\theta(A|X)$ denotes the conditional probability of $y \in A$ given X for a value θ and we denote it $P_0(A|X)$ when $\theta = \theta_0$, the true unknown value. Also, $P_0(X)$ denotes the vector of conditional probability of the r outcomes of $\mathcal{Y}$. We define the identified set for θ, Θ_I, as the collection of parameters which are observationally equivalent to the true value θ_0, i.e.,

$$\Theta_I = \left\{ \theta \in \mathbb{R}^{q+q'}, \forall y \in \mathcal{Y}, P_\theta(Y = y|X) = P_0(Y = y|X),\ X\ a.s. \right\}.$$

Observe that Θ_I might be a true set or a point. Our procedure does not depend on the true nature of the set. If Θ_I were a point, we could imagine using standard procedures to estimate it. However, ignoring some of the inequalities would result in getting larger confidence regions for

[10]In our example, $\mathcal{Y} = \{(0,0), (1,0), \ldots, (3,3)\}$, i.e., 16 possible outcomes.

the parameter of interest θ. In the next part, we show how to collect the set of inequalities which characterize the identified "set".

3.1 A sharp characterization of the identified set

Following Proposition 1, for any $x \in \mathcal{X}$, an outcome $y \in \mathcal{Y}$ is a possible equilibrium of the game if and only if the unobserved shock lies in $\mathcal{R}_\omega(x, y)$.

Given that there exist multiple equilibria regions, the probability of ε to be in $\mathcal{R}_\omega(x, y)$ is an upper bound of the probability to observe the outcome y. Therefore, we get

$$\forall x \in \mathcal{X},\ \forall y \in \mathcal{Y},\ P_0(Y = y|X = x) \leq \int_{\mathcal{R}_\omega(x,y)} dF_\eta(\varepsilon). \tag{2}$$

The inequality in (2) can be extended to any subset $A \subset \mathcal{Y}$ and following the propositions in Beresteanu et al. (2011) and Galichon and Henry (2011) in particular, we can characterize the identified set by a countable number of conditional moment inequalities:

$$\Theta_I = \{\theta \in \Theta \mid \forall A \in \mathcal{Y}, P_0(Y \in A|X) \leq P_\eta(\varepsilon \in \mathcal{R}_\omega(X, A))\ X\ a.s.\}. \tag{3}$$

We call $P_0(Y \in A|X) \leq P_\eta(\varepsilon \in \mathcal{R}_\omega(X, A))$ the (conditional moment) inequality generated by A. Unfortunately, this characterization of the identified set is hardly exploitable as such by the econometrician. Even for the simplest games, the number of inequalities characterizing the sharp identified set can already be overwhelming. For example, when $\mathcal{X}$ is finite, the number of inequalities is $\text{card}(\mathcal{P}(\mathcal{Y})) \times \text{card}(\mathcal{X})$. In the the pedagogical example that we consider and if we assume that the exogenous cost shifter is degenerate ($\text{card}(\mathcal{X}) = 1$), the number of inequalities is $\text{card}(\mathcal{P}(\mathcal{Y})) = 2^{16} = 65536$ which is tractable. With 5 players and three types, a case compatible with the retail or the airline industry, the number of inequalities increases to 2^{125} which is untractable.

Furthermore, with the exception of simple cases, it is not clear how to compute the theoretical bounds implied by the model without making use of simulation methods which necessarily induce biases. For the sake of clarity, let us abstract from the additional layer of difficulty induced by the exogenous covariates X and perform the identification analysis as if we condition on a given realization $X = x$. We postpone the discussion on the inclusion of the covariates to the section on estimation. Accordingly, to lighten the notations, we omit the X in the rest of this section and the inequalities have to be interpreted conditionally on X, X almost surely. We now propose simple solutions to substantially simplify the characterization of the identified set.

3.2 Selection of the Inequalities

The set of inequalities in (3) characterizes sharply the identified set. Any parameter in the identified set satisfies these inequalities and reversely. However,there are two issues.

First, in order to be implementable, one needs to derive the regions $\mathcal{R}_\omega(A)$ for any set of outcomes $\{y_1,\ldots,y_k\}$. Though Proposition 1 gives closed form expressions when A is a single outcome, calculating the regions $R_\omega(A)$ for subsets A composed by several outcomes requires to know whether the different regions $R_\omega(y)$ for $y\in A$ are multiple equilibria regions and how they overlap with other multiple equilibria regions. In other words, characterizing the set of inequalities require to know the structure of the multiple equilibria regions.

Second, the number of inequalities in 3 can be very large. Even in cases where brute force would be possible, the inference procedure would be challenging, in particular to get competitive critical values. Luckily, a lot of these inequalities are redundant in the sense that they are implied by the knowledge of other inequalities. Notice that it is a different notion than the redundancy used in the GMM literature.[11] To be more specific, we provide a definition to be self-contained.

As exhibited previously,

Definition 1 (Redundancy of a moment inequality) *Let $A\in\mathcal{P}(\mathcal{Y})$, we say that A generates a redundant inequality if it exist $A_1\in\mathcal{P}(\mathcal{Y})$ and $A_2\in\mathcal{P}(\mathcal{Y})$ not empty such that*

$$P_0(Y\in A_i)\le P_\eta(\varepsilon\in\mathcal{R}_\omega(A_i)), i=1,2\Rightarrow P_0(Y\in A)\le P_\eta(\varepsilon\in\mathcal{R}_\omega(A)).$$

For example, in our example, $A=\{(1,0),(2,1)\}$ generates a redundant inequality because these two outcomes do not have regions $R_\omega(y)$ which overlap, so

$$R_\omega(A)=R_\omega((1,0))\cup R_\omega((2,1)).$$

Therefore if $P_0(Y=(1,0))\le P_\eta(\varepsilon\in\mathcal{R}_\omega(1,0))$ and $P_0(Y=(2,1))\le P_\eta(\varepsilon\in\mathcal{R}_\omega(2,1))$,

$$P_0(Y\in A)=P_0(Y=(1,0))+P_0(Y=(2,1))\le P_\eta(\varepsilon\in\mathcal{R}_\omega(1,0))+P_\eta(\varepsilon\in\mathcal{R}_\omega(2,1))=P_\eta(\varepsilon\in\mathcal{R}_\omega(A)).$$

[11]See, for example, Breusch et al. (1999).

We propose an algorithm which determines the multiple equilibria structure for a given parameter θ. This algorithm allows us in the same procedure to eliminate most of the redundant inequalities and to calculate the probabilities of the different regions $R_\omega(A),\ A \in \mathcal{P}(\mathcal{Y})$.

The graph generated by the model Before detailing our algorithm, we explain how we eliminate redundant inequalities by sufficient conditions. It has been explained in the literature (Galichon and Henry (2011), Beresteanu et al. (2011), Chesher and Rosen (2017), Bontemps and Kumar (2020), Luo and Wang (2017)) that eliminating redundant inequalities, or, equivalently, determining a core determining class is linked to the graph structure of the model. First, a core determining class of the model is a subset of $\mathcal{P}(\mathcal{Y})$ such that it is sufficient to test the inequalities generates by the elements of this class to have all the inequalities satisfied. When one eliminates redundant moment inequalities, one determine a non trivial core determining class (or set).

The graph $\Gamma(\theta)$ generated by the model is defined as a graph linking the outcomes $y_i \in \mathcal{Y}$ such that there exists an edge between two elements y_1 and y_2 if their equilibrium regions $\mathcal{R}_\omega(y_1)$ and $\mathcal{R}_\omega(y_2)$ overlap. Following the terminology used in graph theory, a subset A is connected in the graph $\Gamma(\theta)$ if and only if there exists a path between every pair of elements in A. The structure of the graph helps us in eliminating redundant inequalities.

Proposition 2 (Sufficient condition for redundancy) *If a subset $A \subset \mathcal{Y}$ is not connected in the graph $\Gamma(\theta)$, then A yields a redundant inequality.*

The proof is simple and therefore omitted see the example above.

In our trade-off between feasibility and efficiency, we decide to eliminate the most possible redundant inequalities to get the smaller set of inequalities to test. Observe however, that adding redundant inequalities might improve the small sample properties of the estimated identified set (in terms of volume for example). In determining a core determining class, i.e., a subset of $\mathcal{P}(\mathcal{Y})$ which yields to a sharp characterization of Θ_I in (3), it is important to remind that it is not unique and can vary in size depending on the conditions that are used to eliminate inequalities. Luo and Wang (2017) provides conditions to find the smallest core determining class in the context of entry games. Bontemps and Kumar (2020) propose a characterization of the smallest core determining class. Our contribution in this paper is to provide a simple construction to derive a core determining class of inequalities by directly deriving the graph $\Gamma(\theta)$.

3.3 Our algorithm to determine a core determining class

In this part, we show how our algorithm allows us to eliminate moment inequalities and to calculate the upper bound of $P(Y \in A)$ for each selected element A of $\mathcal{P}(\mathcal{Y})$. Remark that, in some special case like in Mazzeo (2002) and Cleeren et al. (2010), it is possible to predetermine the graph $\Gamma(\theta)$. However, the restrictions imposed on the two-type model of Cleeren et al. (2010) are difficult to generalize to more types without assuming much stronger restrictions, especially on the horizontal differentiation between the different types.

We now state a necessary and sufficient condition for two equilibrium regions $\mathcal{R}_\omega(y_1)$ and $\mathcal{R}_\omega(y_2)$ to overlap.

Proposition 3 (Overlapping equilibrium regions) *Two outcomes $y = (N_1, ..., N_T)$ and $\bar{y} = (\bar{N}_1, ..., \bar{N}_T)$ have their equilibrium regions $\mathcal{R}_\omega(y)$ and $\mathcal{R}_\omega(\bar{y})$ which overlap if and only if $\forall t \in \mathcal{T}$,*[12]

$$\max\left(-\pi_t(N_t, \mathbf{N}_{-t}; \omega), -\pi_t(\bar{N}_t, \bar{\mathbf{N}}_{-t}; \omega)\right) < \min\left(-\pi_t(N_t + 1, \mathbf{N}_{-t}; \omega), -\pi_t(\bar{N}_t + 1, \bar{\mathbf{N}}_{-t}; \omega)\right) \tag{4}$$

The proof is straightforward given the fact that the regions $\mathcal{R}_\omega(y)$ are cubes in $\mathbb{R}^T$. It can be generalized to any set of outcomes.

Proposition 4 (Characterization of the intersection regions) *For any element A of $\mathcal{P}(\mathcal{Y})$, when it is non-empty, the intersection region $\underset{y_k \in A}{\cap} \mathcal{R}_\omega(y_k)$ is defined in each dimension t in $\mathcal{T}$ as follows:*

$$\max_{y_k \in A} - \pi_t(y_k; \omega) \leq \varepsilon_t < \min_{y_k \in A} - \pi_t(y_k^+; \omega),$$

with $y_k = (N_{t,k}, \mathbf{N}_{-t,k})$ and $y_k^+ = (N_{t,k} + 1, \mathbf{N}_{-t,k})$.

Our algorithm is executed as follows.

Stage 1: Compute the regions $\mathcal{R}_\omega(y)$ for all single outcomes of $\mathcal{Y}$. Collect the moment inequalities generated from (2).

[12] Remember that, by convention, we have that $\forall t \in \mathcal{T}$, $\pi_t(0, \mathbf{N}_{-t}; \omega) = -\infty$ and $\lim_{N_t \to +\infty} \pi_t(N_t, \mathbf{N}_{-t}; \omega) = +\infty$.

Stage 2: Check all the pairs ($K = 2$) to see if the two outcomes y_1 and y_2 of each pair have their associated equilibrium regions $\mathcal{R}_\omega(y_1)$ and $\mathcal{R}_\omega(y_2)$ which overlap using condition (4). Draw an edge between these outcomes (for the graph $\Gamma(\theta)$) if it is the case.

For the pairs $A = \{y_1, y_2\}$ which are connected by an edge, compute the sharp upper bound of $P(Y \in A)$:

$$P(Y \in A) \leq P_\eta(\varepsilon \in \mathcal{R}_\omega(y_1)) + P_\eta(\varepsilon \in \mathcal{R}_\omega(y_2)) - P_\eta(\varepsilon \in \mathcal{R}_\omega(y_1) \cap \mathcal{R}_\omega(y_2)).$$

Add these moment inequalities to the set of inequalities generated by the single outcomes considered in Stage 1.

Stage 3: $K = 3$. We now check all triplets (y_1, y_2, y_3) given that their equilibrium regions might overlap if and only if they overlap two by two. In other words, we focus on the connected pairs to select our "triplet candidates". Again, if the three regions overlap we have a connected subset of three elements, otherwise we eliminate the triplet and do not consider the moment inequality generated by an eliminated triplet.

For the connected subsets of three elements $A = \{y_1, y_2, y_3\}$, compute the sharp upper bound of $P(Y \in A)$:

$$\begin{aligned} P(Y \in A) \leq \; & P_\eta(\varepsilon \in \mathcal{R}_\omega(y_1)) + P_\eta(\varepsilon \in \mathcal{R}_\omega(y_2)) + P_\eta(\varepsilon \in \mathcal{R}_\omega(y_3)) \\ & - P_\eta(\varepsilon \in \mathcal{R}_\omega(y_1) \cap \mathcal{R}_\omega(y_2)) - P_\eta(\varepsilon \in \mathcal{R}_\omega(y_1) \cap \mathcal{R}_\omega(y_3)) - P_\eta(\varepsilon \in \mathcal{R}_\omega(y_2) \cap \mathcal{R}_\omega(y_3)) \\ & + P_\eta(\varepsilon \in \mathcal{R}_\omega(y_1) \cap \mathcal{R}_\omega(y_2) \cap \mathcal{R}_\omega(y_3)). \end{aligned}$$

Stage 4: $K = 4$. Check now for all connected subsets of four elements given that each subset of three elements must be in the remaining connected subsets of three elements and so forth

The algorithm stops when there are no longer connected subsets of $K = l + 1$ elements. Remark that it allows to derive all the components H_i of $\Gamma(\theta)$.[13] For each component, we have a moment equality. Keeping it as an inequality is however sufficient to have a sharp characterization of the set. Given that the region corresponding to the outcome $(0, 0)$ does not overlap with any other region $\mathcal{R}_\omega(y)$, we have at least two components.

[13] The components of a graph are subgraphs $\{H_i\}_{i=1}^k$ such that each H_i is connected and H_i is not connected to H_j for $i \neq j$.

The algorithm stops within a finite number of steps, each of them being polynomial in the cardinality of $\mathcal{Y}$, given the fact that profit of a type t firm is strictly decreasing in the number of active firms of this type and tends to $-\infty$ when N_t tends to infinity. When a subset $A = \{y_1, \ldots, y_K\}$ of K elements is selected by the algorithm, the sharp upper bound for the probability of $Y \in A$ can be derived using the expression:

$$
\begin{aligned}
P_\eta(\varepsilon \in \mathcal{R}_\omega(A)) &= P_\eta(\varepsilon \in \underset{y_i \in A}{\cup} \mathcal{R}_\omega(y_i)) \\
&= \sum_{k=1}^{K} P_\eta(\varepsilon \in \mathcal{R}_\omega(y_k)) + \sum_{i=2}^{K} (-1)^{i-1} \sum_{A_i \subset A \ \& \, \mathrm{card}(A_i) = i} P_\eta(\varepsilon \in \underset{y_k \in A_i}{\cap} \mathcal{R}_\omega(y_k)).
\end{aligned}
$$

If this subset is a component of the graph $\Gamma(\theta)$, it generates an equality. An important remark here is that these intersection regions are also cubes in $\mathbb{R}^T$ and thus integrating over these regions is straightforward.

Application to our example Figure 3 displays the resulting graph for our example. In this case there are no connected subsets of three elements. The identified set can therefore be characterized sharply by 14 inequalities related to the single outcomes which are not $(0, 0)$ and $(3, 3)$, 9 inequalities generated by the selected pairs and 7 equalities generated by the 7 components (including $(0, 0)$ and $(3, 3)$).

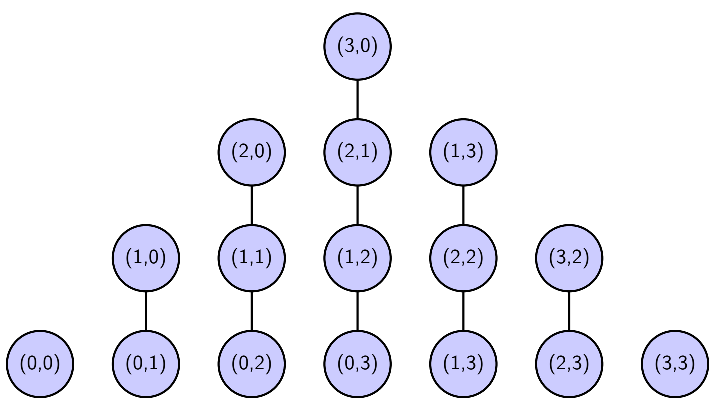


Figure 3: $\Gamma(\theta)$ for $\beta_1 = 3$, $\beta_2 = 2$, $\delta_{11} = \delta_{22} = 1.5$ and $\delta_{12} = \delta_{21} = 0.5$

Remark on the algorithm

- The cost to check whether a subset of p elements is connected is low because it is just comparing the maximum of p quantities with the minimum of p other quantities following Proposition 4.

- The collection of all connected subsets constitutes a core determining class $\mathcal{C}(\theta)$, i.e. they generate moment inequalities which characterize sharply the identified set Θ_I. However, it might not be the smallest class.

- One advantage of our sequential algorithm is that it allows the researcher to stop the procedure after a few iterations. This can prove useful in practice if the set of connected subsets is too large and it also allows to measure the effect on the identified set of going one iteration further in the collection of connected subsets.

- If one wants to compute the minimum probability of any subset, one need to execute the algorithm until the end to determine all the possible multiplicities. This is the reason why Ciliberto and Tamer (2009) derives the lower bound for each probability of a single outcome using simulation methods. However, simulating these bounds introduce noises which are often (wrongly) ignored in the inference procedure.

4 Inference on the full vector

The last section allows us to test that parameter θ is in in the identified set Θ_I from moment inequalities generated by a subset of $\mathcal{P}(\mathcal{Y})$, called $\mathcal{C}(\theta, X)$ which may depend on X. Calling p_X the number of elements of $\mathcal{C}(\theta, X)$, the moment inequality generated by any member A_j of $\mathcal{C}(\theta, X)$ can be rewritten in a simple form:

$$q_j^\top P_0(X) \leq C_{\theta,j}(X),$$

in which q_j is a $r \times 1$ vector of zeros and ones, the ones corresponding to the outcomes present in A_j and $C_{\theta,j}(X) = P_\eta(\varepsilon \in \mathcal{R}_\omega(X, A_j))$.

Calling D_X the stacked version of the $q_j^\top$s and $C_\theta(X)$ the stacked version of the $C_{\theta,j}(X)$s, we get:

$$\begin{aligned} \theta \in \Theta_I &\iff D_X P_0(X) \leq C_\theta(X),\ X\ a.s. \\ &\iff \mathbb{E}\left(C_{\theta,j}(X) - q_j^\top \mathbf{1}(Y|X)\right) \geq 0, \forall\ j = 1, \ldots, p_X\ X\ a.s. \end{aligned} \tag{5}$$

We have a collection of conditional moment inequalities $m_j(Y, X, \theta)$ which are linear in $\mathbf{1}(Y|X)$.

Moreover, for a given $X = x$, the condition $D_X P \leq C_\theta(x)$ in (5) defines a convex set $A(\theta, x)$ in which $P_0(x)$ should lie for θ to be in the identified set (see Figure 4).

Remark that when we turn to inference, only $P_0(X)$ is estimated as we invert a test (θ is fixed). The critical part is to define a procedure which remains simple to implement, because we have to repeat it for each candidate θ, but sufficiently powerful to avoid estimating too large confidence regions. In this section, we propose a new method which seems to satisfy these requirements. Additionally, explanatory variables create a curse of dimensionality when they are taken into account. In most of the cases, researchers have discretized them but at the cost of changing the original model. Here, our method can handle non-parametric estimation of $P_0(X)$ with a slight correction of the test statistic.

In the following, we assume we observe a i.i.d. sample of n outcomes $(X_1, Y_1), \ldots, (X_n, Y_n)$ in independent markets. We first study the inference procedure in the absence of covariates and we show how to adapt the procedures to include them.

4.1 Inference without covariates

Without covariayes, we can drop in this part the dependance of the structural quantities in X. Let P_n be the empirical frequency vector of outcomes:

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(y_i = Y),$$

in which the inequality $\mathbf{1}(y_i = Y)$ should be interpreted term by term, i.e.,

$$\mathbf{1}(y_i = Y) = \left[\mathbf{1}(y_i = y_1), \mathbf{1}(y_i = y_2), \ldots, \mathbf{1}(y_i = y_{card\mathcal{Y}})\right]^\top.$$

Our empirical analogue of the moment inequalities is:

$$\bar{m}_j(Y, \theta) = \frac{1}{n} \sum_{i=1}^{n} m_j(y_i, \theta) = C_{\theta,j} - q_j^\top P_n, \ j = 1, \ldots, p$$

which should be the estimates of positive quantities. Also, we denote $\Sigma_n = diag(P_n) - P_n^\top P_n$, i.e., a consistent estimator of $Var P_n$.[14]

[14]Under the i.i.d. assumption, the frequency estimator is asymptotically normal and

$$\sqrt{n}(P_n - P_0) \xrightarrow[n \to \infty]{d} N(0, diag(P_0) - P_0 P_0^\top).$$

4.1.1 Existing procedures

Various test statistics have been proposed in the literature to test the moment inequality restrictions in Equation (5). Andrews and Soares (2010) propose different test statistics and the GMS procedure to calculate accurate critical values. Andrews and Barwick (2012) study the performance of these different values and provide guidance about the tuning parameters involved in the procedure. In a more recent work, Chernozhukov et al. (2018b) propose a test statistic easy to compute with a critical value which is valid whatever the correlation structure of the moments involved.[15] Their approach is particularly suited for cases such as games, which display a very large number of inequalities. Contrary to alternatives such as subsampling or general moment selection, the critical value, which is based on a moderate deviation inequality for self-normalized sums, is straightforward to compute and increases (in absolute value) slowly with the number of moments. This property is particularly attractive as we need to repeat the testing procedure for each point in the grid but still be competitive while using the procedure.

Minimum test statistic To be more specific, let $\xi_n(\theta)$ be defined as the minimum over the studentized moments:

$$\xi_n(\theta) = \min_{j=1,\dots,p} \frac{\sqrt{n}(C_{\theta,j} - q_j^\top P_n)}{\sqrt{q_j^\top \Sigma_n q_j}},$$

When θ does not belong to the identified set, the quantity above should diverge to $-\infty$. When $\theta \in \Theta_I$, the asymptotic distribution of $\xi_n(\theta)$ can be derived[16] and it is equal to

$$\min_{j\in\mathcal{J}(\theta)} \frac{q_j^\top Z}{\sqrt{q_j^\top \Sigma_0 q_j}},$$

in which Z follows a normal distribution with variance Σ_0 and $\mathcal{J}(\theta)$ is the collection of indices j corresponding to the binding moments. This asymptotic distribution depends on the number and the identity of the binding moments, as expected. In the following, p^* denotes the number of binding moments, i.e., the cardinal of $\mathcal{J}(\theta)$.

[15]For papers applying this procedure, see for instance Bontemps and Kumar (2020), Chesher and Rosen (2019).
[16]See Bontemps and Kumar (2020), Proposition 9.

A critical value can be computed after a first step estimation of the set of binding moments $\mathcal{J}(\theta)$ like in the GMS procedure of Andrews and Soares (2010). Simulation methods (bootstrap and/or subsampling techniques) can be also considered to improve the accuracy of the critical value. Chernozhukov et al. (2018b) propose the following one:

$$c^*(\alpha) = \frac{\Phi^{-1}(\alpha/p)}{\sqrt{1 - \Phi^{-1}(\alpha/p)^2/n}} \tag{6}$$

where $\Phi(\cdot)$ is the c.d.f. of the standard normal distribution ($\Phi^{-1}(\cdot)$ is its inverse). The advantage of this critical value is that it is easy to compute, quite competitive and it decreases at a rate of the order $-\sqrt{\log(p/\alpha)}$, i.e., do not diverge too quickly when the number of moments is high. Under some mild regularity assumptions, satisfied in our framework, Chernozhukov et al. (2018b) show[17] that the confidence set $\mathrm{CR}_n(1-\alpha)$ induced by $c^*(\alpha)$

$$\mathrm{CR}_n(1-\alpha) = \{\theta \in \Theta \mid \xi_n(\theta) \geq c^*(\alpha)\}$$

is asymptotically valid, i.e.,

$$\liminf_{n\to\infty} \inf_{\theta\in\Theta_I} \Pr(\theta \in CR_n) \geq 1 - \alpha.$$

Using the convex set theory Alternatively, one can exploit the equivalence in (5) between testing the moment inequalities and testing that P_0 belongs to the convex set $A(\theta)$. Convexity is an attractive feature which have been exploited in the set identification literature by Beresteanu et al. (2011) or by Bontemps and Kumar (2020) for games with multiple equilibria. In particular, the authors use the support function

$$\delta^*(q; A(\theta)) = \sup_{P\in A(\theta)} q^\top P.$$

Following Rockafellar (1997),

$$P_0 \in A(\theta) \Longleftrightarrow \min_q \delta^*(q; A(\theta)) - q^\top P_0 \geq 0.$$

In other words, the support function embeds all the moment inequalities.

[17] See Theorem 4.1

Classical reformulation of the problem above shows that the program is strictly equivalent to test the (euclidian) distance between P_0 and $A(\theta)$. Therefore, we can consider test statistic based on generalized distance:

$$d_\Omega(P_0, A(\theta)) = \min_{P, DP \le C_\theta} (P_0 - P)^\top \Omega^{-1} (P_0 - P). \tag{7}$$

As $A(\theta)$ is a convex set, the distance can be easily computed from quadratic solvers under linear constraints. However, the asymptotic distribution of $d_\Omega(P_n, A(\theta))$ is still a complicated distribution because it depends on whether the true value P_0 is inside the set, is a vertex, lies on an exposed face or an edge of the convex set. One can work with conservative critical values which are easy to compute like the critical value proposed by Cox and Shi (2022) for a conditional version of the test or use the old literature on inequality testing and the upper bound proposed by some of the authors (see Silvapulle and Sen, 2011 for a review of the existing procedures). Kitamura and Stoye (2018) propose simulation methods which are valid in our framework.

In the next section, we propose a procedure which is asymptotically pivotal and, therefore, avoids the difficulty of the derivation of the critical value.

4.1.2 The smoothed approach

As illustrated above, the difficulty in the traditional moment inequality approach is to recover the exact asymptotic distribution of the test statistic, which depends on the set of binding moments which defines whether the true vector $P_0(Y)$ lies inside, on an exposed face of a given order or is a vertex. While computationally intensive methods [18] often better approximate the exact asymptotic distribution, the implementation difficulties make these methods unappealing for the estimation of games. On the other hand, the usage of upper bounds on the asymptotic or exact distribution of the test statistic results in conservative confidence regions.

Here, we propose to approximate $A(\theta)$ by a smooth outer set. The advantage of manipulating a smooth set is that we can recover a test statistic with a pivotal asymptotic distribution[19] due to the fact that only one equality is binding. This approach completely removes the inferential

[18] Even if they don't eliminate the inferential loss.

[19] In particular, it doesn't involve any convoluted and numerically-intensive inference procedure such as bootstrap or sub-sampling to derive the critical value.

loss caused by the inability to derive the exact asymptotic distribution by introducing a small and manageable identification loss through smoothing, which we make vanish asymptotically. The resulting confidence regions have the right size asymptotically. Last but not least, we see in the next section how this smoothing procedure facilitates the introduction of covariates in the model.

The log-sum exponential function Let us now introduce our new estimation strategy. From what precedes, the sharp identified is such that:

$$\begin{aligned} \theta \in \Theta_I &\iff \min_{j=1,\dots,p} \mathbb{E} m_j(Y,\theta) \geq 0, \\ &\iff P_0 \in A(\theta). \end{aligned} \tag{8}$$

In order to smooth the identified set, we suggest to replace the minimum operator by a smooth approximation. For $z = (z_1, z_2, ..., z_p) \in \mathbb{R}^p$, we have that a smooth approximation of the minimum function writes:

$$g_\rho(z) = -\rho^{-1} \log\Big(\sum_{j=1}^{p} \exp(-\rho z_j)\Big),$$

in which ρ, the smoothing parameter, controls the level of approximation. $g_\rho(\cdot)$ is known as the log-sum exponential (LSE) function used in machine-learning and numerical optimization.

A nice property of this approximation is that it is possible to control for the difference between the minimum and its approximation through the following inequality:

$$0 \leq \min_{1\leq j\leq p} z_j - g_\rho(z) \leq \rho^{-1}\log(p). \tag{9}$$

This inequality is straightforward to derive and the upper bound is reached when all the elements in z are equal. The minimum is simply the limit of $g_\rho(z)$ when $\rho \to +\infty$.

A smooth outer set Let $A^o_\rho(\theta)$ defined as:

$$A^o_\rho(\theta) = \left\{ P \in \mathbb{R}^T \mid G_{\theta,\rho}(P) = -\rho^{-1}\log\left(\frac{1+\sum_{j=1}^p \exp(-\rho(C_{\theta,j} - q_j^\top P))}{p+1}\right) \geq 0 \right\}. \tag{10}$$

We know collect the set $\Theta^o_I(\rho)$ of parameters θ such that

$$\begin{aligned} \Theta^o_I(\rho) &= \left\{\theta \in \mathbb{R}^{q+q'} \mid P_0 \in A^o_\rho(\theta)\right\}, \\ &= \left\{ \theta \in \mathbb{R}^{q+q'} \mid G_{\theta,\rho}(P_0) = -\rho^{-1}\log\left(\frac{1+\sum_{j=1}^p \exp(-\rho(C_{\theta,j} - q_j^\top P_0))}{p+1}\right) \geq 0 \right\}. \end{aligned} \tag{11}$$

The next proposition shows that it is a smooth outer approximation of the true identified set Θ_I.

Proposition 5 *The following statements hold*

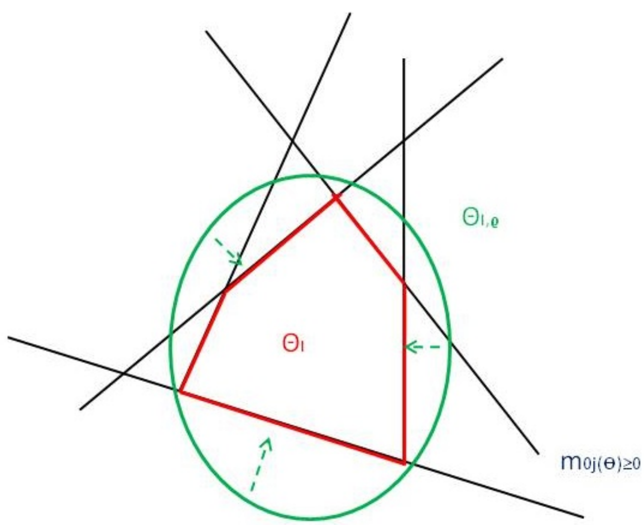
(i) For any $\rho > 0$, $\Theta_I \subset \Theta_I^o(\rho)$

(ii) $A_\rho^o(\theta)$ *is a strictly convex set and, therefore, has no exposed face and no kink.*

(iii) $\lim_{\rho \to +\infty} d_H(\Theta_I, \Theta_I^o(\rho)) = 0$, *where* d_H *is the Hausdorff distance.*

See the proof in the Appendix.

Figure 4: Smoothed outer set with different values of ρ



Remark that the strict convexity of the set $A_\rho^o(\theta)$ comes from the addition of the quantity 1 inside the sum in the logarithm. As $\min_{j=1,\ldots,p} \mathbb{E}m_j(Y,\theta) \geq 0$ is equivalent to $\min\left(0, \min_{j=1,\ldots,p} \mathbb{E}m_j(Y,\theta)\right) = 0$, adding $e^0 = 1$ in the LSE function does not change the strategy, but makes the smooth approximation of the minimum strictly convex.[20]

[20]Otherwise, let $e = (1, 1 \ldots, 1) \in \mathbb{R}^p$ and $\lambda \in \mathbb{R}$, $g_\rho(\lambda e) = \log p/\rho + \lambda$ which is a linear function.

A pivotal test statistic As the function $-G_{\theta,\rho}(\cdot)$ is strictly convex and infinitely differentiable in a neighborhood of P_0, we can build a regular test statistic. Let us consider the empirical analogue of $G_{\theta,\rho}(P_0)$, in which we let ρ depending on the sample size n and in which we plug P_n in replacement of P_0. Under the null, $G_{\theta,\rho_n}(P_n)$ should be the estimate of a positive quantity. The following proposition shows how to recenter $\sqrt{n}G_{\theta,\rho_n}(P_n)$ to get an asymptotically normal distribution.

Proposition 6 *Let* $\nabla G_{\theta,\rho}(P) = -\frac{\sum_{j=1}^{p} q_j \exp(-\rho(C_{\theta,j}-q_j^\top P))}{1+\sum_{j=1}^{p}\exp(-\rho(C_{\theta,j}-q_j^\top P))}$, $\Sigma_0 = diag(P_0) - P_0 P_0^\top$ *and* $\Gamma_0(\theta) = -\frac{\sum_{j\in\mathcal{J}(\theta)} q_j}{1+p^*}$. *Let* ρ_n *a divergent sequence of positive number such that* $\rho_n = O\left(\frac{1}{n^a}\right)$, $0 < a < 1/2$. *Then,*

$$\sqrt{n}\left(G_{\theta,\rho_n}(P_n) - G_{\theta,\rho_n}(P_0)\right) \xrightarrow[n\to\infty]{d} N(0, \Gamma_0(\theta)\Sigma_0\Gamma_0(\theta)^\top). \tag{12}$$

The proof is in the appendix and exploits the fact that $G_{\theta,\rho}(\cdot)$ is twice-continuously differentiable. The constraint on the speed of ρ controls for the higher order terms. Finally, $\Gamma_0(\theta)$ is the limit of $\nabla G_{\theta,\rho}(P_0)$ when ρ tends to infinity.

Wecon Observe that Proposition 6 requires the knowledge of the binding moments to derive $\Gamma_0(\theta)$ and to calculate $G_{\theta,\rho_n}(P_0)$. We now make it applicable. First, we define $V_n = \nabla G_{\theta,\rho_n}(P_n)\Sigma_n \nabla G_{\theta,\rho_n}(P_n)^\top$. We show it is a consistent estimator of the variance of the asymptotic distribution in (12).[21] Then, we estimate the number of binding moments p^* by collecting the number of empirical moments which are "close to 0", i.e., for τ_n such that τ_n tends to 0 and $\sqrt{n}\tau_n$ tends to $+\infty$,

$$\widehat{p^*}_n = \sum_{j=1}^{p} \mathbf{1}\left\{\frac{C_{\theta,j} - q_j^\top P_n}{\sqrt{q_j^\top \Sigma_n q_j}} < \tau_n\right\}. \tag{13}$$

The following proposition holds.

Proposition 7 *Let* $\rho_n = cn^\alpha$ *with* $0 < \alpha < \frac{1}{2}$ *and* $c > 0$. *Let* $\widehat{p^*}_n$ *an estimator of the number of binding moments,* p^**, defined in Equation (13). Then, we have under* H_0*:*

$$\sqrt{n}\frac{G_{\theta,\rho_n}(P_n) - \frac{1}{\rho_n}\log\left(\frac{1+p}{1+\widehat{p^*}_n}\right)}{\sqrt{V_n}} \xrightarrow[n\to\infty]{d} N(0,1). \tag{14}$$

[21] See Lemma 2 in the Appendix.

Once, the asymptotic distribution of the recentered test statistic has been derived, we show that our procedure gives valid confidence region with asymptotically exact size. We consider two different constructions. In the first one, $\mathrm{CR}_n(1-\alpha)$ is the region defined by

$$\mathrm{CR}_n(1-\alpha) = \{\theta \in \mathbb{R}^{q+q'} \ , \sqrt{n}\frac{G_{\theta,\rho_n}(P_n) - \frac{1}{\rho_n}\log\left(\frac{1+p}{1+\widehat{p^*}_n}\right)}{\sqrt{V_n}} \geq z_\alpha\}, \tag{15}$$

in which z_α is the α-quantile of the standard normal distribution. In the second one, we ignore the recentering term and consider the region $\widetilde{\mathrm{CR}}_n(1-\alpha)$ defined by

$$\widetilde{\mathrm{CR}}_n(1-\alpha) = \{\theta \in \mathbb{R}^{q+q'} \ , \sqrt{n}\frac{G_{\theta,\rho_n}(P_n)}{\sqrt{V_n}} \geq z_\alpha\}, \tag{16}$$

Then, the following proposition holds

Proposition 8 (Validity and consistency of the confidence regions) *The confidence regions $\mathrm{CR}_n(1-\alpha)$ and $\widetilde{\mathrm{CR}}_n(1-\alpha)$ defined in Equation (15) and (16) are asymptotically valid and consistent, i.e.,*

- *Asymptotic validity:* $\liminf_{n\to\infty} \inf_{\theta\in\Theta_I} P_0(\theta \in \mathrm{CR}_n(1-\alpha)) \geq 1-\alpha$ *and* $\liminf_{n\to\infty} \inf_{\theta\in\Theta_I} P_0(\theta \in \widetilde{\mathrm{CR}}_n(1-\alpha)) \geq 1-\alpha$.
- *Consistency:* $\forall \theta \notin \Theta_I$, $P_0(\theta \in \mathrm{CR}_n(1-\alpha)) \underset{n\to\infty}{\longrightarrow} 0$ *and* $P_0(\theta \in \widetilde{\mathrm{CR}}_n(1-\alpha)) \underset{n\to\infty}{\longrightarrow} 0$.

Therefore, our procedure is quite simple. The test statistic and the critical value are straightforward to derive. Remark that the exponential in the expression makes the non binding moments vanishing in the numerical evaluation of $G_{\theta,\rho_n}(P_n)$ when ρ_n is large. Also, it is obvious to see that

$$\mathrm{CR}_n(1-\alpha) \subset \widetilde{\mathrm{CR}}_n(1-\alpha).$$

However, $\widetilde{\mathrm{CR}}_n(1-\alpha)$ does not require to estimate p^*.

Remark

Given the expression of V_n above, it is always possible to calculate it, *ex-ante*, and to evaluate the following quantity instead of $G_{\theta,\rho_n}(P_n)$:

$$\tilde{G}_{\theta,\rho_n}(P_n) = -\rho_n^{-1}\log\left(\frac{1+\sum_{j=1}^{p}\exp\left(-\frac{\rho_n(C_{\theta,j}-q_j^\top P_n)}{\sqrt{V_n}}\right)}{p+1}\right),$$

i.e. to standardize the moment $m_j(\cdot)$ inside the log-sum exponential function. The advantage is to have a procedure which is robust to the unit of measurement of the moments m_js. This alternative is evaluated in the Monte Carlo section.

4.2 Inference with covariates

The introduction of covariates doesn't substantially modify the derivation of the sharp identified set for the structural parameter θ. In particular, the algorithm to select the inequalities and the approach to compute the theoretical bounds developed in section 3.3 remain the same but have to be performed for each element $x \in \mathcal{X}$. However, the estimation of the conditional probabilities of the different outcomes is done at a non regular rate and should change the derivations done in the previous subsection. From now on, we denote by $h_0(X)$ the vector of conditional probabilities (instead of $P_0(X)$). Following (5), we recall an equivalent characterization of the identifies set Θ_I in the presence of covariates:

$$\begin{aligned}\theta \in \Theta_I &\Longleftrightarrow h_0(X) \in A(\theta, X)\ X\ a.s.\\ &\Longleftrightarrow D_X h_0(X) \leq C_\theta(X),\ X\ a.s.\\ &\Longleftrightarrow \mathbb{E}\left(m_j(Y, X, \theta)|X\right) = \left(C_{\theta,j}(X) - q_j^\top h_0(X)\right) \geq 0, \forall\ j = 1, \ldots, p_X\ X\ a.s.\end{aligned}$$

A priori, the sharp characterization of the convex set $A(\theta, X)$ depends on X. Deriving the core determining class for any value of x can be challenging when the dimension of X increases. Also, repeating the algorithm introduced in section 3.3 may be computationally expensive. To keep things tractable numerically, we make a separable assumption for our profit functions:

Assumption 3 (Additively separable profit shifters)

$$\pi_t(X, N_t, \mathbf{N}_{-t}; \omega) = \kappa_t(X; \omega_1) + \phi_t(N_t, \mathbf{N}_{-t}; \omega_2), \quad \forall t = 1, \ldots, T.$$

The next proposition states that, with the profit structure assumed in Assumption 3, the core determining class remains the same for all $x \in \mathcal{X}$.

Proposition 9 (Invariant core determining class) *Under Assumptions 1, 2, 3, we have:*

$$\forall x \in \mathcal{X}, \quad \mathcal{C}(\theta, x) = \mathcal{C}(\theta).$$

It implies that under assumption 3, the core determining class only needs to be computed once for each candidate θ (as opposed to deriving the core determining for every combination of x and θ). The intuition for this result is that under additive separability of the covariates in the profit functions, the covariates only translates the equilibrium structure in the space of unobserved shocks and the graph remains the same.

From now on, we assume Assumption 3 and we denote the core determining class $\mathcal{C}(\theta)$ since it does not depend on X. p is its cardinality.

4.2.1 From conditional moment inequalities to unconditional moment inequalities

Following (5), we have p conditional moment inequalities which should be positive almost surely in X. Conditional moment inequalities are much more difficult to tackle than unconditional ones. First, given that X is continuous, the set is characterized by an infinite number of (unconditional) inequalities. Second, conditional moments are non-parametric objects that are harder to estimate than unconditional moments and which display non-standard asymptotic properties. Various methods have been proposed to estimate confidence sets in models defined by conditional moment inequalities (Andrews and Shi (2013), Armstrong and Chan (2016), Armstrong (2014) among others) and theoretical econometricians have suggested to directly apply these methods to the case of entry games.

The leading method (proposed in Andrews and Shi (2013)) consists in transforming the conditional moment inequalities into unconditional ones. More precisely, Andrews and Shi (2013) considers a collection $\mathcal{G}$ of non-negative functions of X, denoted $g(X)$. These functions allow us to transform the conditional moment inequalities into unconditional ones as follows:

$$m_{j,g}(Y,X,\theta) = \left(C_{\theta,j}(X) - q_j^\top \mathbf{1}(Y)\right) g(X)$$

and the resulting outer set that is characterized by the unconditional moment inequalities writes:

$$\bar{\Theta}_I = \{\theta \in \Theta \mid \mathbb{E}\left[m_{j,g}(Y,X,\theta)\right] \geq 0,\ \forall j \in \{1,\ldots,p\},\ \forall g \in \mathcal{G}\}.$$

Under high level conditions on $\mathcal{G}$, the outer set defined above coincides with the sharp identified set. The choice of $\mathcal{G}$ is critical: they suggest to use a countable family of hypercubes. For the estimation, they integrate these unconditional moments into either a Cramer Von Mises (CvM) or a Kolmogorov Smirnov (KS) type of statistic. Finally, they adapt the GMS procedure to derive the critical value.

The inference strategy in Andrews and Shi (2013) is extremely challenging numerically with a lot of tuning parameters (choice of integration functions, derivation of the critical value...). To our knowledge, a few papers only used it for real empirical applications.

Researchers often favor an approach based on discretizing the support of continuous variables (even if it results in a modification of the initial model and thus of the identified set) or picking particular $g(\cdot)$ functions in $\mathcal{G}$ above. For example, Aradillas-Lopez and Rosen (2022) use the density of X as weighting function.

4.2.2 Sharp characterization of the identified set in the case with covariates

Let us define the following moment, which serves as the basis of our estimation procedure:

$$m(X, \theta) = \min_{j=1,\dots,p} \left\{ C_{\theta,j}(X) - q_j^\top h_0(X) \right\}.$$

The sharp identified set can easily be expressed in terms of the previous moment, from a choice of $g(\cdot)$ which is positive, smooth, and does not vanish on the support of X.

Proposition 10

$$\begin{aligned} \theta \in \Theta_I &\iff m(X, \theta) = 0 \ X \ a.s \\ &\iff \mathbb{E}[m(X, \theta) g(X)] = 0. \end{aligned} \tag{17}$$

First, the minimum of $m_j(X, \theta)$ is zero and not weakly positive because we exploit the structure of the game which provides equalities (i.e., one inequality and its "opposite" inequality) due to the presence of components in the graph $\Gamma(\theta, X)$, in particular the ones related to the outcome "No firm enters" and its complement. As a result, we have at least four binding inequalities (i.e. two equalities) and the minimum is zero.

Then, the second equivalence might be striking at a first glance. It is usually known as an implication. Again, we exploit the fact that there is (at least) two equalities hidden in the list of the p inequalities. If $\theta \notin \Theta_I$, because $\mathcal{C}(\theta)$ is a core determining class, there exist $\eta < 0$, and $j_0 \leq p$

such that:

$$P_X\left(m_{j_0}(x,\theta) < \eta < 0\right) > 0_0.$$

Outside the set of X which satisfy the restriction above, the minimum of the $m_j(x,\theta)$ can not be positive because of the equalities involved in the remaining list. Therefore:

$$\mathbb{E}[m(X,\theta)g(X)] < 0.$$

Proposition 10 is particularly important because it characterize sharply the identified set with one moment inequality. There is no need to consider a countable family of weighting function. In practice, we choose $g = 1$ but we could choose g equal to the density of X.

$m(\cdot,\theta)$ be defined as the minimum of p functions, we can apply the smooth approximation of the minimum function used in the section without covariate.

4.2.3 Smooth test statistic

Let us slightly modify the definition of the smoothing function $G_{\theta,\rho}$ introduced above. We make the dependence in $h_0(X)$ explicit because we have to treat the non-parametric rate of convergence of its estimator $\hat{h}_0(X)$.

$$\begin{aligned} G_{\theta,\rho}: \ & \mathcal{X} \times \mathcal{H} \to \mathbb{R} \\ & (x,h) \mapsto -\rho^{-1} \log\left(\frac{1 + \sum_{j=1}^{p} \exp(-\rho(C_{\theta,j}(X) - q_j^\top h_0(X)))}{1+p}\right). \end{aligned}$$

We have the following implications:[22]

$$\theta \in \Theta_I \implies G_{\theta,\rho}(X,h_0) \geq 0, \quad X\text{a.s} \implies \mathbb{E}\left[G_{\theta,\rho}(X,h_0)\right] \geq 0.$$

Like before, our idea is to test the last inequality after having plugged a consistent estimator of $h_0(X)$ and let ρ diverging at an appropriate rate. However, the expansion in Equation (12) in the non covariate case relies on the square-root convergence of P_n to h_0 in the absence of covariances. Here, the non-parametric rate affects the derivation of the asymptotic distribution. This type of the problem has been extensively studied in the literature on semi-parametric estimation (Newey (1994),

[22]In the expression $G_{\theta,\rho}(X,h_0)$ is short form of $G_{\theta,\rho}(X,h_0(X))$.

Ai and Chen (2003), Chernozhukov et al. (2018a), Ackerberg et al. (2014)) and a key regularity condition to recover a square-root asymptotic distribution is for the moment to be "differentiable" with respect to the first stage non-parametric estimator. We start by deriving the asymptotic distribution of $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} G_{\theta,\rho}(X_i,\hat{h}_0) - \mathbb{E}\left[G_{\theta,\rho}(X,h_0)\right]\right)$. In particular, we exploit some existing results in the literature on semi-parametric estimation (Ackerberg et al. (2014) and Newey (1994)). Before we state the result, we list some regularity conditions on the non-parametric estimator that we require to derive the asymptotic distribution.

Assumption 4 *(Regularity conditions for non-parametric estimator).*

1. $\sup\limits_{x\in\mathcal{X}} |\hat{h}_0(x) - h_0(x)| = O_p(n^{-\gamma})$ *with* $\gamma > 1/4$.
2. *Stochastic equicontinuity:*
$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial G_{\theta,\rho}(X_i,h_0)'}{\partial h}(\hat{h}_0(X_i) - h_0(X_i)) - \int \frac{\partial G_{\rho}(x,h_0)'}{\partial h}(\hat{h}_0(x) - h_0(x))\mathrm{d}x\right] \xrightarrow{P} 0.$$

We consider kernel based non-parametric estimator for h_0 and we make following assumptions on smoothness of kernel as well as the function itself. We now state additional regularity assumptions on the kernel estimator to prove mean square differentiability and stochastic equicontinuity .

Assumption 5 *The supp$(X) = \mathcal{X}$ is compact and $f_0(\cdot)$, the p.d.f. of X is bounded away from zero as well as bounded above.*

1. *The kernel $K(\cdot)$ is differentiable of order β with bounded derivatives, $K(\cdot)$ is zero outside the bounded set, $\int K(u)du = 1$ and there is positive integer m such that for $j < m$,*
$$\int u^j K(u)du = 0.$$
2. *The density $f_0(\cdot)$ and regression function $h_0(\cdot)$ both are continuously differentiable of order d with bounded derivatives in an open set containing $\mathcal{X}$.*

In (i), we require that the rate of convergence of $\hat{h}_0$ be faster than $n^{-1/4}$ to ensure that higher order terms in the expansion vanish when n goes to ∞. This rate can be achieved by most non-parametric estimators under appropriate smoothness conditions on h_0. Stochastic equicontinuity is a well known condition in the semi-parametric literature. In the same spirit as Newey (1994), we provide lower level conditions on the kernel estimator under which (i) and (ii) are satisfied for the kernel estimator. Under assumptions 4, we are able to derive the following linear expansion.

Proposition 11 (Second order expansion with covariates) *Let $W_i = (X_i, Y_i)$. Under assumption 4, there exists $\tilde{h} \in \mathcal{H}$, such that*

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} G_{\theta,\rho}(X_i,\hat{h}_0) - \mathbb{E}\left[G_{\theta,\rho}(X_i,h_0)\right]\right) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} G_{\theta,\rho}(X_i,h_0) + \alpha(W_i) - \mathbb{E}\left[G_{\theta,\rho}(X_i,h_0)\right]\right)$$
$$+ \frac{\rho}{\sqrt{n}}\sum_{i=1}^{n}(\hat{h}_0(X_i) - h_0(X_i))^\top J_\rho(X_i,\tilde{h})(\hat{h}_0(X_i) - h_0(X_i)) + o_p(1),$$

with $\alpha(W_i) = \frac{\partial G_{\theta,\rho}(X_i,h_0)}{\partial h}^\top(\mathbf{1}\{Y_i = y\} - h_0(X_i))$ and for any $(x_i,\tilde{h})$, $||J_\rho(x_i,\tilde{h})||_\infty \leq 2p$.

Several remarks are in order. First, $\alpha(\cdot)$ is the adjustment term that arises because of the first stage estimation of h_0. Second, the 2nd order term in the expansion is of order $O_p(\rho n^{1/2-2\gamma})$ and thus, constrains the rate of divergence of ρ toward $+\infty$ in order this term to go to 0 asymptotically. It is linked to the non-parametric convergence rate of $\hat{h}_0$, $\gamma > 1/4$.

Building on proposition 11, it is straightforward to construct a valid and consistent confidence region for Θ_I.

Proposition 12 *Let $\rho_n = cn^\alpha$ with $\alpha < 2\gamma - \frac{1}{2}$ and $c > 0$ a constant, let V_n an estimator of the asymptotic variance of $G_{\theta,\rho}(X_i,h_0) + \alpha(W_i)$, let $\hat{h}_0$ a non-parametric estimator satisfying Assumptions 4 and 5. Let the confidence region $\mathrm{CR}_n(1-\alpha)$ defined by:*

$$\mathrm{CR}_n(1-\alpha) = \{\theta \in \Theta \mid \sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^{n} G_{\theta,\rho}(X_i,\hat{h}_0)}{\sqrt{V_n}} \geq z_\alpha\}.$$

Then, $\mathrm{CR}_n(1-\alpha)$ is asymptotically valid and consistent, i.e.,

- *Asymptotic validity:* $\liminf_{n\to\infty} \inf_{\theta\in\Theta_I} \Pr(\theta \in \mathrm{CR}_n(1-\alpha)) \geq 1-\alpha$.
- *Consistency:* $\forall\theta \notin \Theta_I$, $\Pr(\theta \in \mathrm{CR}_n(1-\alpha)) \to 0$.

See the proof in the appendix. Proposition 12 mimicks Proposition 8 with an additional term in the variance due to the nonparametric estimation of h_0. The rate of growth of ρ_n is however more restricted. The same results hold if we use a series estimator rather than kernel estimator for h_0.

4.2.4 Debiasing Approach

Instead of correcting the variance due to the estimation of h_0, we can change the function we use in order to be invariant to the true estimator used for $\hat{h}_0$. Following Ackerberg et al. (2014), we can write orthogonal moments as

$$\tilde{G}_{\theta,\rho}(X_i, h_0) = G_{\theta,\rho}(X_i, h_0) - \frac{\partial G_{\theta,\rho}(X_i, h_0)}{\partial h}^{\top} (\mathbf{1}\{Y_i = y\} - h_0(X_i)).$$

We introduce additional parameter α as

$$\mathbb{E}[\tilde{G}_{\theta,\rho}(X_i, h_0)] - \alpha = 0,$$

and our goal is to find an estimator of α along with its asymptotic distribution. Following Chernozhukov et al. (2018a), we split the sample into K equal samples $S_1, \ldots, S_K$. We define our estimator as

$$\hat{\alpha} = \frac{1}{K} \sum_{i=1}^{K} \frac{1}{n_k} \sum_{j \in S_i} \tilde{G}_{\theta,\rho}(X_j, \hat{h}_{-i}(X_j)),$$

where h_{-i} is an estimator of h calculated using sample S_{-i}. We just need to verify assumption 3.1 and 3.2 of Theorem 3.1 in Chernozhukov et al. (2018a) in order to show

$$\sqrt{n}(\hat{\alpha} - \alpha_0) \xrightarrow{d} \mathcal{N}(0, \Omega_0),$$

where

$$\Omega_0 = \mathbb{E}\left[\left\{G_\rho(X_i) - \frac{\partial G_{\theta,\rho}(X_i)}{\partial h}^{\top} (\mathbf{1}\{Y_i = y\} - h_0(X_i))\right\}\left\{G_\rho(X_i) - \frac{\partial G_{\theta,\rho}(X_i)}{\partial h}^{\top} (\mathbf{1}\{Y_i = y\} - h_0(X_i))\right\}^{\top}\right].$$

We provide a more general version of debiasing where we also allow the smoothing parameter ρ to increase to infinite as sample size increase.

Proposition 13 (Debiasing with decreasing smoothing) *If $\rho_n = cn^{\alpha}$ with $\alpha < 2\gamma - \frac{1}{2}$ and c a constant, then*

$$\sqrt{n}\left(\hat{\alpha} - \alpha\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\tilde{G}_{\theta,\rho}(X_i, h_0) - \mathbb{E}(\tilde{G}_{\theta,\rho}(X_i, h_0))\right] + o\left(1\right).$$

5 Monte Carlo simulations

We perform Monte Carlo simulations to evaluate the relative performance of the different procedures proposed in this paper of the identified set as well as the different estimation procedures.

5.1 Simulations without covariates

Simulation Design The simulation design directly follows the example introduced in this paper for different sample sizes. Profit functions of firms of type 1 and 2 write as follows:

$$\Pi_1 = \beta_1 - \delta_{1,1}N_1 - \delta_{2,1}N_2 + \varepsilon_1$$

$$\Pi_2 = \beta_2 - \delta_{1,2}N_1 - \delta_{2,2}N_2 + \varepsilon_2,$$

in which

- N_i is the number of firms of type i=1,2
- ε_i for $i = 1, 2$ is the profit shocks. We assume that they are i.i.d., each of them drawn from a standard distribution

In our DGP, we have $\beta_1 = 3$, $\beta_2 = 2$, $\delta_{11} = \delta_{22} = 1.5$ and $\delta_{12} = \delta_{21} = 0.5$. To make the exposition easier, we assume that the econometrician knows β_1 and β_2 and that $\delta_{11} = \delta_{22}$ as well as $\delta_{12} = \delta_{21}$. This is of course, unrealistic but allows us to simplify the exposition of the results. The general case does present the same type of conclusions. In the multiple equilibria regions, we assume that a firm of type 1 always decides first and, therefore, we pick the equilibrium with the highest number of type 1 firms. Observe that following Cleeren et al. (2010), we know that in the multiple equilibria regions, $N_1 + N_2$ is invariant.

The graph related to the equilibrium structure is given in Figure 5. Therefore, we can sharply characterize the identified set from 9 inequalities derived from each single outcome, completed by 4 inequalities derived from pairs of outcomes (the links in the graph) and one inequality related to the component of outcomes with $N_1 + N_2 = 3$, i.e., a total of 14 inequalities. Among these fourteen inequalities, 5 are equalities (we have five components).

The sample size is $n = 1,000$. The number of Monte Carlo replications is 1,000. For each sample, we compute the decision to reject or not $\theta \in \Theta_I$ for a 5% level of significance. The grid tested is composed by values from 1 to 2 with a tick of 0.02 for δ_{ii} and values from 0.4 to 1.4 with a tick of 0.02 for δ_{ij}, i.e., 2601 points tested in total. We report the mean across simulations of the lowest value and highest value for the two parameters tested. We also report the mean number of points not rejected as well as the coverage rate of the true value θ_0. Given the DGP, we can calculate P_0 which is

$$P_0 = [0.021, 0.074, 0.256, 0.047, 0.131, 0.421, 0.012, 0.034, 0.0004]^\top.$$

Different procedures evaluated Following Section 3, we can derive different sets of inequalities to test our candidates θ. We now detail them.

- *core*. Here, we select the set of inequalities which corresponds to the core determining class $\mathcal{C}(\theta)$ that we derived by gathering all the connected subsets as shown in section 3.3, applying our algorithm.

- *core*$^+$. Here, we add to the previous set of inequalities, the five equalities satisfied by the model. It is not necessary. In fact, if $P_1 \leq Q_1$ and $P_2 \leq Q_2$ with $P_1 + P_2 = 1$ and $Q_1 + Q_2 = 1$, then $P_1 = Q_1$ and $P_2 = Q_2$. Nevertheless, we decide to add these five inequalities to reinforce the equality requirement.

- *core**. Here, we add to the previous set of inequalities, four out of the five equalities satisfied by the model. The fifth one be redundant, we wonder what would be the impact of ignoring one of them (we drop the last one related to the outcome $(2, 2)$.

- $\min - \max$ set of inequalities which corresponds to an upper bound and a lower bound on the probability of each individual outcome

- $\min - \max^+$. Here we add the three equalities related to the non-single components to the min max inequalities.

Let us emphasize that for all these sets of inequalities, we derive the exact bounds implied by the model as opposed to the rest of the literature, which except in simple cases, simulates the bounds. Thus, the min-max strategy, which has been exploited in earlier work (like in Ciliberto and Tamer (2009) for example) can already be understood as an enhanced version of the min-max strategy.

Concerning the critical value, we used Equation (6) with the number of inequalities tested (labeled $CCK, p^* =$). In some cases, we were able to refine this maximum number of binding moment inequalities exploiting the geometry of the set $A(\theta)$. We also use Liu and Xie (2020) (label "cauchy") who propose to aggregate the p-values of the different inequalities/equalities tested into one single test statistic. Interestingly, the critical value they propose is valid for any correlation structure.

Finally, we also smooth from each set of inequalities following our methodology and we present the results for different values of ρ.

We display the results in Table 1 and 2. Let us first focus on the first set of results. Except for the case $\min - \max^+$ with the smallest amount of degrees of freedom, the size is controlled for the true value. Following the specific geometry, we know that 10 out of 14 inequalities are binding at the maximum. Given that firm of type 1 always enter first, P_0 is a kink of the convex set $A(\theta)$ and, therefore, these 10 equalities are indeed biding with $\theta = \theta_0$. Interestingly, despite the fact that testing these fourteens inequalities is equivalent to testing these fourteen inequalities completed by the 5 equalities (i.e. adding the "opposite" inequalities for five of them), the small sample properties are much better in the latter case. One of the message is to include these equalities explicitly in the testing procedure. Deleting one of the redundant one ($core^*$ $CCK,$ p^* =13) does not help.

The min max procedure seems competitive. We recall that computing the minimum of the probability of a single outcome requires to run our algorithm to the end, which could require a lot of numerical evaluations in more complicated DGPs, or to simulate it (like in Ciliberto and Tamer, 2009) and incorporate an additional noise in the testing procedure. Adding equalities to the min-max procedure seems to work well.

Finally, the aggregation of all test into one dilutes power and does not provide a competitive alternative.

In Table 2, we evaluate the same set of equaliteis/inequalities with our smooth min function $G_{\theta,\rho}(\cdot)$ with three values of ρ, 1, 5 and 10. Again, for these procedures, the test statistic is asymptotically pivotal. We consider two versions, one in which we compte the moments $C_{\theta,j} - q_j^\top P_n$ for each j, one in which we standardize each of these moments before incorporating them in the calculation of the test statistic. First, as expected, higher values of ρ lead to smaller but valid confidence regions, which are competitive with respect to the best procedures of Table 1. Second, it is better to normalize before aggregating, because it avoids the results to be driven by the moment of the highest variance.

We change the DGP for the results displayed in Table 3 and 4 by considering uniform profit shocks instead of normally distributed ones. The impact on the data is to get smaller multiple equilibria regions.[23] The results remain qualitatively the same. And so, for the same DGP for which we allocate randomly the outcome in the multiple equilibria regions (Table 5 and 6).

5.2 Simulations with covariates

We know consider the case with covariates, in a similar setting than above. Here, X follows uniform distribution on $[0, 1]$. Given X the profit of both types is equal to:

$$\Pi_1 = \beta_1 + \beta_X X - \delta_{1,1} N_1 - \delta_{2,1} N_2 + \varepsilon_1$$
$$\Pi_2 = \beta_2 + \beta_X X - \delta_{1,2} N_1 - \delta_{2,2} N_2 + \varepsilon_2,$$

In our DGP, we have $\beta_1 = 3$, $\beta_2 = 2$, $\delta_{11} = \delta_{22} = 1.5$ and $\delta_{12} = \delta_{21} = 0.5$, like before, and $\beta_X = 1$. In a first step, the shocks are normally distributed like before.

We compare two strategies. Remember that for our specification, the core determining class does not depend on the realization of X. As a result, one necessary condition for θ to be in the identified set is that the unconditional probability vector (estimated at the standard rate) should belong to a convex set which is the Aumann expectation of $A(\theta, X)$ (see Beresteanu and Molinari (2008)). Therefore, after having calculated the expectation of the quantities involved in the case without covariate (with respect to X), we can perform the same procedure than before in Table 1. Results are displayed in Table 7. Alternatively, we can apply our smooth procedure and the results are displayed in Table 8.

First, the size properties are less good than in the case without covariates. When ρ is too large, the smooth "approximation" is not competitive. It is the case whether we use a point, i.e.

[23]For the values chosen, the total area of multiplicity is around 6% against 15% for the DGP with normal shocks. Here the new equilibrium probabilities are

$$P_0 = [0.104, 0.078, 0.135, 0.151, 0.047, 0.333, 0.036, 0.094, 0.021]^\top.$$

$\mathbb{E}_X h_0(X)$ or our procedure with $h_0(X)$. The same hierarchy than without the covariates still holds. The results seem to depend on the "size" of the multiple equilibria regions. They are better with the uniform shocks, see Table 9. Alternative estimators for the variance calculation of the quantity V_n in 12 should be perhaps considered. It is left for future research.

6 Conclusion

In this paper, we develop a new methodology to estimate entry games with multiple equilibria, which may or may be not point identified. First, we propose an algorithm which allows us to characterize the equilibrium structure in polynomial time (in the number of types). This algorithm permits to derive a competitive core determining class and to calculate the bounds used for the moment inequalities generated by this class.

Then, we propose to circumvent the problem of deriving a competitive critical value but computable in a reasonable amount of time by smoothing the minimum function. The smooth min or max functions have been used in applied mathematics. In our case, it allows us to obtain a pivotal test statistic which automatically eliminates "numerically" the non binding moments. Obviously, there is a trade-off between high values of the smoothing parameter to get close to the true identified set and accuracy of the normal approximation. Values of $\rho = 5$ seem to be competitive though they should be confirmed with more simulations.

Interestingly, our procedure can easily be adapted to the case with covariates, either by testing the unconditional probability vector or by adapting our asymptotic distribution to a non parametric estimator of the conditional probability vector. The square-root speed of convergence of our statistic is recovered. Monte Carlo simulations, study the properties of our procedure and underline the fact that adding redundant moments in the procedure may improve the small sample properties and that size properties of our test are very sensitive to the plug-in of the empirical variance.

Many pending questions remain. First, a more precise guideline about the choice of the smoothing parameter ρ could be investigated. Then,inference on subvectors have been proposed in very general settings. It would be worth investigating whether the specific structure of entry games would allow an improvement in the application of these techniques. Finally, our methodology could be evaluated on real data.

Tables of results

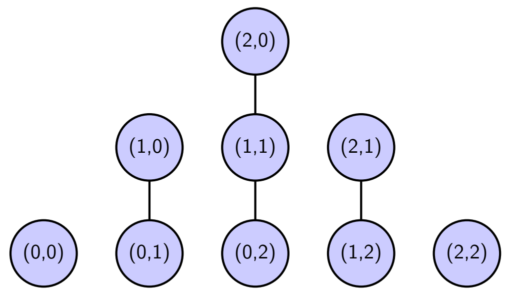


Figure 5: $\Gamma(\theta)$ for $\beta_1 = 3$, $\beta_2 = 2$, $\delta_{11} = \delta_{22} = 1.5$ and $\delta_{12} = \delta_{21} = 0.75$

Table 1: Coverage rate and confidence region - normal shocks

	Coverage	$\min \delta_{11}$	$\max \delta_{11}$	$\min \delta_{12}$	$\max \delta_{12}$	Nb. points
core $CCK,\ p^* = 10$	0.98	1.39	1.58	0.55	1.00	133
core $CCK,\ p^* = 14$	0.99	1.38	1.58	0.54	1.02	148
core$^+$ $CCK,\ p^* = 15$	0.95	1.41	1.57	0.63	0.99	105
core* $CCK,\ p^* = 13$	0.95	1.41	1.57	0.62	1.04	123
core Cauchy	1.00	1.16	1.75	0.41	1.40	976
$\min - \max$ $CCK,\ p^* = 11$	0.90	1.38	1.60	0.63	1.02	139
$\min - \max$ $CCK,\ p^* = 18$	0.94	1.38	1.61	0.62	1.05	163
$\min - \max$ Cauchy	1.00	1.00	2.00	0.40	1.40	2257
$\min - \max^+$ $CCK,\ p^* = 17$	0.89	1.41	1.57	0.63	0.97	97
$\min - \max^+$ $CCK,\ p^* = 24$	0.93	1.41	1.58	0.63	0.99	108

$\beta_1 = 3$, $\beta_2 = 2$, $\delta_{11} = \delta_{22} = 1.5$ and $\delta_{12} = \delta_{21} = 0.75$

Table 2: Coverage rate and confidence region - normal shocks - Pivotal tests

	Coverage	$\min \delta_{11}$	$\max \delta_{11}$	$\min \delta_{12}$	$\max \delta_{12}$	Nb. points
Smooth-*core* NS $\rho = 1$	1.00	1.00	1.73	0.40	1.40	1716
Smooth-*core* NS $\rho = 5$	1.00	1.08	1.65	0.43	1.40	1170
Smooth-*core* NS $\rho = 10$	1.00	1.25	1.62	0.50	1.40	719
Smooth-*core* STD $\rho = 1$	1.00	1.00	1.82	0.40	1.40	1811
Smooth-*core* STD $\rho = 5$	1.00	1.24	1.64	0.41	1.31	667
Smooth-*core* STD $\rho = 10$	1.00	1.34	1.60	0.51	1.07	273
Smooth-$core^+$ NS $\rho = 1$	1.00	1.00	1.72	0.40	1.40	1659
Smooth-$core^+$ NS $\rho = 5$	1.00	1.19	1.64	0.44	1.40	912
Smooth-$core^+$ NS $\rho = 10$	1.00	1.31	1.61	0.52	1.40	561
Smooth-$core^+$ STD $\rho = 1$	1.00	1.00	1.75	0.40	1.40	1440
Smooth-$core^+$ STD $\rho = 5$	1.00	1.30	1.61	0.54	1.32	421
Smooth-$core^+$ STD $\rho = 10$	1.00	1.37	1.58	0.61	1.06	171
Smooth-$\min - \max$ NS $\rho = 1$	1.00	1.00	2.00	0.40	1.40	2547
Smooth-$\min - \max$ NS $\rho = 5$	1.00	1.01	1.85	0.45	1.40	1749
Smooth-$\min - \max$ NS $\rho = 10$	1.00	1.19	1.75	0.51	1.40	1114
Smooth-$\min - \max$ STD $\rho = 1$	1.00	1.00	1.99	0.43	1.40	2014
Smooth-$\min - \max$ STD $\rho = 5$	1.00	1.23	1.69	0.57	1.40	825
Smooth-$\min - \max$ STD $\rho = 10$	0.99	1.33	1.62	0.62	1.26	380
Smooth-$\min - \max^+$ NS $\rho = 1$	1.00	1.00	1.97	0.40	1.40	2237
Smooth-$\min - \max^+$ NS $\rho = 5$	1.00	1.20	1.68	0.48	1.40	1031
Smooth-$\min - \max^+$ NS $\rho = 10$	1.00	1.31	1.63	0.54	1.40	647
Smooth-$\min - \max^+$ STD $\rho = 1$	1.00	1.00	1.82	0.45	1.40	1645
Smooth-$\min - \max^+$ STD $\rho = 5$	1.00	1.31	1.63	0.58	1.39	500
Smooth-$\min - \max^+$ STD $\rho = 10$	0.98	1.38	1.59	0.63	1.10	188

$\beta_1 = 3$, $\beta_2 = 2$, $\delta_{11} = \delta_{22} = 1.5$ and $\delta_{12} = \delta_{21} = 0.75$

Table 3: Coverage rate and confidence region - unif. shocks

	Coverage	$\min \delta_{11}$	$\max \delta_{11}$	$\min \delta_{12}$	$\max \delta_{12}$	Nb. points
core CCK, $p^* = 10$	0.98	1.29	1.72	0.47	1.08	398
core CCK, $p^* = 14$	0.99	1.28	1.74	0.45	1.10	448
$core^+$ CCK, $p^* = 15$	0.95	1.31	1.68	0.56	0.99	231
$core^*$ CCK, $p^* = 13$	0.96	1.23	1.68	0.46	1.13	517
core Cauchy	0.98	1.01	1.94	0.45	1.40	1273
$\min - \max$ CCK, $p^* = 11$	0.94	1.29	1.68	0.56	1.04	248
$\min - \max$ CCK, $p^* = 18$	0.95	1.27	1.70	0.54	1.07	303
$\min - \max$ Cauchy	0.99	1.00	2.00	0.40	1.40	2028
$\min - \max^+$ CCK, $p^* = 17$	0.93	1.32	1.67	0.56	0.99	221
$\min - \max^+$ CCK, $p^* = 24$	0.94	1.30	1.68	0.55	1.01	251

$\beta_1 = 3$, $\beta_2 = 2$, $\delta_{11} = \delta_{22} = 1.5$ and $\delta_{12} = \delta_{21} = 0.75$

Table 4: Coverage rate and confidence region - unif. shocks - Pivotal tests

	Coverage	$\min \delta_{11}$	$\max \delta_{11}$	$\min \delta_{12}$	$\max \delta_{12}$	Nb. points
Smooth-*core* NS ρ =1	1.00	1.00	2.00	0.54	1.40	2063
Smooth-*core* NS ρ =5	1.00	1.00	1.90	0.56	1.40	1753
Smooth-*core* NS ρ =10	0.99	1.03	1.79	0.58	1.40	1418
Smooth-*core* STD ρ =1	1.00	1.00	1.93	0.40	1.40	2369
Smooth-*core* STD ρ =5	1.00	1.12	1.73	0.47	1.30	1026
Smooth-*core* STD ρ =10	1.00	1.25	1.67	0.54	1.08	459
Smooth-$core^+$ NS ρ =1	1.00	1.00	1.99	0.55	1.40	2012
Smooth-$core^+$ NS ρ =5	1.00	1.01	1.83	0.57	1.40	1554
Smooth-$core^+$ NS ρ =10	0.99	1.16	1.74	0.59	1.37	954
Smooth-$core^+$ STD ρ =1	1.00	1.00	1.91	0.45	1.40	1943
Smooth-$core^+$ STD ρ =5	1.00	1.20	1.71	0.56	1.26	681
Smooth-$core^+$ STD ρ =10	0.96	1.30	1.64	0.61	1.05	291
Smooth-$\min - \max$ NS ρ =1	1.00	1.00	2.00	0.40	1.40	2597
Smooth-$\min - \max$ NS ρ =5	1.00	1.00	2.00	0.42	1.40	2269
Smooth-$\min - \max$ NS ρ =10	1.00	1.05	1.98	0.47	1.40	1788
Smooth-$\min - \max$ STD ρ =1	1.00	1.00	2.00	0.40	1.40	2293
Smooth-$\min - \max$ STD ρ =5	1.00	1.10	1.80	0.53	1.40	1253
Smooth-$\min - \max$ STD ρ =10	0.99	1.25	1.69	0.59	1.23	551
Smooth-$\min - \max^+$ NS ρ =1	1.00	1.00	2.00	0.40	1.40	2563
Smooth-$\min - \max^+$ NS ρ =5	1.00	1.05	1.93	0.46	1.40	1793
Smooth-$\min - \max^+$ NS ρ =10	1.00	1.18	1.78	0.51	1.37	1077
Smooth-$\min - \max^+$ STD ρ =1	1.00	1.00	2.00	0.42	1.40	2189
Smooth-$\min - \max^+$ STD ρ =5	1.00	1.20	1.73	0.56	1.30	781
Smooth-$\min - \max^+$ STD ρ =10	0.97	1.31	1.65	0.61	1.07	321

$\beta_1 = 3$, $\beta_2 = 2$, $\delta_{11} = \delta_{22} = 1.5$ and $\delta_{12} = \delta_{21} = 0.75$

Table 5: Coverage rate and confidence region - unif. shocks -interior point

	Coverage	$\min \delta_{11}$	$\max \delta_{11}$	$\min \delta_{12}$	$\max \delta_{12}$	Nb. points
core $CCK,\ p^* =10$	0.99	1.29	1.73	0.46	1.08	408
core $CCK,\ p^* =14$	0.99	1.28	1.74	0.45	1.10	458
core$^+$ $CCK,\ p^* =15$	0.96	1.31	1.68	0.55	0.99	234
core* $CCK,\ p^* =13$	0.97	1.21	1.68	0.45	1.16	551
core Cauchy	0.98	1.01	1.92	0.46	1.40	1150
$\min - \max$ $CCK,\ p^* =11$	0.96	1.28	1.69	0.54	1.05	277
$\min - \max$ $CCK,\ p^* =18$	0.97	1.26	1.71	0.52	1.08	335
$\min - \max$ Cauchy	0.99	1.00	2.00	0.40	1.40	2002
$\min - \max^+$ $CCK,\ p^* =17$	0.96	1.31	1.67	0.55	0.99	235
$\min - \max^+$ $CCK,\ p^* =24$	0.97	1.30	1.68	0.54	1.01	267
Smooth-*core* NS $\rho =1$	1.00	1.00	2.00	0.54	1.40	2062
Smooth-*core* NS $\rho =5$	1.00	1.00	1.90	0.56	1.40	1758
Smooth-*core* NS $\rho =10$	1.00	1.03	1.79	0.57	1.40	1433
Smooth-*core* STD $\rho =1$	1.00	1.00	1.92	0.40	1.40	2318
Smooth-*core* STD $\rho =5$	1.00	1.12	1.72	0.47	1.30	992
Smooth-*core* STD $\rho =10$	1.00	1.25	1.66	0.54	1.08	460
Smooth-*core*$^+$ NS $\rho =1$	1.00	1.00	1.99	0.55	1.40	2010
Smooth-*core*$^+$ NS $\rho =5$	1.00	1.01	1.83	0.57	1.40	1559
Smooth-*core*$^+$ NS $\rho =10$	0.99	1.16	1.74	0.58	1.37	967
Smooth-*core*$^+$ STD $\rho =1$	1.00	1.00	1.90	0.47	1.40	1879
Smooth-*core*$^+$ STD $\rho =5$	0.99	1.20	1.70	0.57	1.25	658
Smooth-*core*$^+$ STD $\rho =10$	0.96	1.30	1.64	0.61	1.05	289
Smooth-$\min - \max$ NS $\rho =1$	1.00	1.00	2.00	0.40	1.40	2598
Smooth-$\min - \max$ NS $\rho =5$	1.00	1.00	2.00	0.41	1.40	2283
Smooth-$\min - \max$ NS $\rho =10$	1.00	1.05	1.98	0.46	1.40	1816
Smooth-$\min - \max$ STD $\rho =1$	1.00	1.00	2.00	0.40	1.40	2298
Smooth-$\min - \max$ STD $\rho =5$	1.00	1.09	1.80	0.52	1.40	1284
Smooth-$\min - \max$ STD $\rho =10$	1.00	1.23	1.69	0.58	1.24	597
Smooth-$\min - \max^+$ NS $\rho =1$	1.00	1.00	2.00	0.40	1.40	2564
Smooth-$\min - \max^+$ NS $\rho =5$	1.00	1.04	1.93	0.45	1.40	1810
Smooth-$\min - \max^+$ NS $\rho =10$	1.00	1.17	1.78	0.51	1.38	1101
Smooth-$\min - \max^+$ STD $\rho =1$	1.00	1.00	2.00	0.42	1.40	2192
Smooth-$\min - \max^+$ STD $\rho =5$	1.00	1.20	1.73	0.55	1.31	806
Smooth-$\min - \max^+$ STD $\rho =10$	0.99	1.30	1.66	0.60	1.08	347

$\beta_1 = 3$, $\beta_2 = 2$, $\delta_{11} = \delta_{22} = 1.5$ and $\delta_{12} = \delta_{21} = 0.75$

Table 6: Coverage rate and confidence region - normal. shocks -interior point

	Coverage	$\min\delta_{11}$	$\max\delta_{11}$	$\min\delta_{12}$	$\max\delta_{12}$	Nb. points
core $CCK,\ p^*=$10	0.99	1.39	1.59	0.49	1.01	162
core $CCK,\ p^*=$14	1.00	1.38	1.59	0.49	1.02	178
core$^+$ $CCK,\ p^*=$15	0.95	1.41	1.59	0.60	0.99	115
core* $CCK,\ p^*=$13	0.96	1.40	1.59	0.59	1.07	145
core Cauchy	0.99	1.16	1.78	0.42	1.40	974
$\min-\max$ $CCK,\ p^*=$11	0.94	1.35	1.61	0.60	1.06	201
$\min-\max$ $CCK,\ p^*=$18	0.96	1.35	1.62	0.59	1.09	230
$\min-\max$ Cauchy	1.00	1.00	2.00	0.40	1.40	2193
$\min-\max^+$ $CCK,\ p^*=$17	0.93	1.41	1.59	0.60	0.99	115
$\min-\max^+$ $CCK,\ p^*=$24	0.96	1.41	1.59	0.59	1.00	126
Smooth-*core* NS $\rho=$1	1.00	1.00	1.73	0.40	1.40	1689
Smooth-*core* NS $\rho=$5	1.00	1.09	1.66	0.41	1.40	1165
Smooth-*core* NS $\rho=$10	1.00	1.25	1.63	0.46	1.40	736
Smooth-*core* STD $\rho=$1	1.00	1.00	1.80	0.40	1.40	1736
Smooth-*core* STD $\rho=$5	1.00	1.24	1.65	0.40	1.25	625
Smooth-*core* STD $\rho=$10	1.00	1.34	1.61	0.49	1.05	290
Smooth-*core*$^+$ NS $\rho=$1	1.00	1.00	1.72	0.40	1.40	1631
Smooth-*core*$^+$ NS $\rho=$5	1.00	1.20	1.64	0.43	1.40	905
Smooth-*core*$^+$ NS $\rho=$10	1.00	1.31	1.62	0.49	1.40	566
Smooth-*core*$^+$ STD $\rho=$1	1.00	1.00	1.72	0.40	1.40	1349
Smooth-*core*$^+$ STD $\rho=$5	1.00	1.30	1.61	0.54	1.26	376
Smooth-*core*$^+$ STD $\rho=$10	1.00	1.37	1.59	0.60	1.04	171
Smooth-$\min-\max$ NS $\rho=$1	1.00	1.00	2.00	0.40	1.40	2561
Smooth-$\min-\max$ NS $\rho=$5	1.00	1.02	1.86	0.42	1.40	1789
Smooth-$\min-\max$ NS $\rho=$10	1.00	1.20	1.75	0.47	1.40	1150
Smooth-$\min-\max$ STD $\rho=$1	1.00	1.00	1.99	0.42	1.40	2010
Smooth-$\min-\max$ STD $\rho=$5	1.00	1.22	1.69	0.55	1.40	873
Smooth-$\min-\max$ STD $\rho=$10	1.00	1.31	1.64	0.60	1.27	438
Smooth-$\min-\max^+$ NS $\rho=$1	1.00	1.00	1.97	0.40	1.40	2255
Smooth-$\min-\max^+$ NS $\rho=$5	1.00	1.21	1.69	0.45	1.40	1050
Smooth-$\min-\max^+$ NS $\rho=$10	1.00	1.31	1.64	0.51	1.40	664
Smooth-$\min-\max^+$ STD $\rho=$1	1.00	1.00	1.81	0.44	1.40	1639
Smooth-$\min-\max^+$ STD $\rho=$5	1.00	1.31	1.64	0.57	1.39	530
Smooth-$\min-\max^+$ STD $\rho=$10	0.99	1.37	1.60	0.61	1.13	222

$\beta_1 = 3$, $\beta_2 = 2$, $\delta_{11} = \delta_{22} = 1.5$ and $\delta_{12} = \delta_{21} = 0.75$

Table 7: Coverage rate and confidence region - normal. shocks - with covariates

	Coverage	$\min \delta_{11}$	$\max \delta_{11}$	$\min \delta_{12}$	$\max \delta_{12}$	Nb. points
core $CCK,\ p^* = 10$	0.98	1.37	1.63	0.65	0.95	113.00
core$^+$ $CCK,\ p^* = 15$	0.88	1.37	1.59	0.66	0.90	78.00
core* $CCK,\ p^* = 13$	0.93	1.35	1.59	0.65	0.91	97.00
$\min - \max^+$ $CCK,\ p^* = 17$	0.83	1.38	1.59	0.66	0.89	73.00
$\min - \max^+$ $CCK,\ p^* = 24$	0.85	1.37	1.59	0.65	0.90	81.00
Smooth-*core* P_0 $\rho = 1$	1.00	1.00	1.99	0.40	1.40	2196.00
Smooth-*core* P_0 $\rho = 5$	1.00	1.18	1.70	0.54	1.14	563.00
Smooth-*core* P_0 $\rho = 10$	1.00	1.30	1.63	0.61	0.99	233.00
Smooth-*core*$^+$ P_0 $\rho = 1$	1.00	1.00	1.84	0.42	1.40	1626.00
Smooth-*core*$^+$ P_0 $\rho = 5$	0.90	1.23	1.65	0.60	1.08	343.00
Smooth-*core*$^+$ P_0 $\rho = 10$	0.89	1.33	1.60	0.64	0.96	141.00
Smooth-$\min - \max$ P_0 $\rho = 1$	1.00	1.00	1.98	0.44	1.40	2019.00
Smooth-$\min - \max$ P_0 $\rho = 5$	0.90	1.15	1.68	0.59	1.34	758.00
Smooth-$\min - \max$ P_0 $\rho = 10$	0.90	1.28	1.63	0.64	1.05	264.00
Smooth-$\min - \max^+$ P_0 $\rho = 1$	1.00	1.00	1.85	0.47	1.40	1711.00
Smooth-$\min - \max^+$ P_0 $\rho = 5$	0.90	1.23	1.64	0.61	1.10	364.00
Smooth-$\min - \max^+$ P_0 $\rho = 10$	0.90	1.33	1.60	0.65	0.95	144.00

$\beta_1 = 3$, $\beta_2 = 2$, $\delta_{11} = \delta_{22} = 1.5$ and $\delta_{12} = \delta_{21} = 0.75$

Table 8: Coverage rate and confidence region - normal. shocks -interior point

	Coverage	$\min \delta_{11}$	$\max \delta_{11}$	$\min \delta_{12}$	$\max \delta_{12}$	Nb. points
Smooth-*core* h_0 $\rho = 1$	1.00	1.00	2.00	0.40	1.40	2204.00
Smooth-*core* h_0 $\rho = 5$	0.98	1.35	1.76	0.59	1.35	599.00
Smooth-*core*$^+$ h_0 $\rho = 1$	0.95	1.00	1.89	0.44	1.40	1655.00
Smooth-*core*$^+$ h_0 $\rho = 5$	0.63	1.38	1.70	0.66	1.26	332.00
Smooth-$\min - \max$ h_0 $\rho = 1$	1.00	1.00	1.98	0.48	1.40	1975.00
Smooth-$\min - \max$ h_0 $\rho = 5$	0.89	1.26	1.80	0.64	1.40	830.00
Smooth-$\min - \max^+$ h_0 $\rho = 1$	0.99	1.00	1.92	0.51	1.40	1726.00
Smooth-$\min - \max^+$ h_0 $\rho = 5$	0.67	1.37	1.70	0.68	1.33	406.00

$\beta_1 = 3$, $\beta_2 = 2$, $\delta_{11} = \delta_{22} = 1.5$ and $\delta_{12} = \delta_{21} = 0.75$

Table 9: Coverage rate and confidence region - unif. shocks -interior point

	Coverage	$\min \delta_{11}$	$\max \delta_{11}$	$\min \delta_{12}$	$\max \delta_{12}$	Nb. points
core CCK, $p^*=$15	0.97	1.28	1.70	0.55	0.96	256.00
Minmax CCK, $p^*=$11	0.93	1.25	1.68	0.55	1.01	263.00
Minmax CCK, $p^*=$18	0.96	1.23	1.70	0.53	1.03	308.00
$\min-\max^+$ CCK, $p^*=$17	0.94	1.28	1.68	0.56	0.96	237.00
$\min-\max^+$ CCK, $p^*=$24	0.96	1.27	1.69	0.55	0.97	264.00
Smooth-*core* P_0 $\rho=$1	1.00	1.00	2.00	0.40	1.40	2300.00
Smooth-*core* P_0 $\rho=$5	1.00	1.15	1.84	0.50	1.33	1124.00
Smooth-*core* P_0 $\rho=$10	0.99	1.26	1.72	0.56	1.08	460.00
Smooth-*core*$^+$ P_0 $\rho=$1	1.00	1.00	2.00	0.43	1.40	2122.00
Smooth-*core*$^+$ P_0 $\rho=$5	0.99	1.20	1.76	0.56	1.24	744.00
Smooth-*core*$^+$ P0y $\rho=$10	0.98	1.30	1.67	0.61	1.03	304.00
Smooth-$\min-\max$ P_0 $\rho=$1	1.00	1.00	2.00	0.40	1.40	2366.00
Smooth-$\min-\max$ P_0 $\rho=$5	1.00	1.07	1.81	0.52	1.39	1308.00
Smooth-$\min-\max$ P_0 $\rho=$10	0.99	1.23	1.69	0.59	1.12	490.00
Smooth-$\min-\max^+$ P_0-rho=y1	1.00	1.00	2.00	0.42	1.40	2288.00
Smooth-$\min-\max^+$ P_0 $\rho=$5	1.00	1.17	1.74	0.55	1.24	792.00
Smooth-$\min-\max^+$ P_0 $\rho=$10	0.98	1.29	1.66	0.61	1.02	317.00
Smooth-*core* h_0 $\rho=$1	1.00	1.00	2.00	0.40	1.40	2290.00
Smooth-*core* h_0 $\rho=$5	1.00	1.19	1.95	0.48	1.39	1434.00
Smooth-*core* h_0 $\rho=$10	0.90	1.33	1.84	0.56	1.21	673.00
Smooth-*core*$^+$ h_0 $\rho=$1	1.00	1.00	2.00	0.41	1.40	2134.00
Smooth-*core*$^+$ h0y $\rho=$5	0.97	1.23	1.88	0.54	1.36	1055.00
Smooth-*core*$^+$ h0y $\rho=$10	0.71	1.37	1.78	0.61	1.15	445.00
Smooth-$\min-\max$ h_0 $\rho=$1	1.00	1.00	2.00	0.41	1.40	2325.00
Smooth-$\min-\max$ h_0 $\rho=$5	1.00	1.12	1.93	0.55	1.40	1441.00
Smooth-$\min-\max$ h_0 $\rho=$10	0.85	1.31	1.80	0.63	1.30	691.00
Smooth-$\min-\max^+$ h_0 $\rho=$1	1.00	1.00	2.00	0.42	1.40	2259.00
Smooth-$\min-\max^+$ h_0 $\rho=$5	0.98	1.22	1.85	0.58	1.36	1024.00
Smooth-$\min-\max^+$ h_0 $\rho=$10	0.74	1.36	1.75	0.65	1.15	414.00

$\beta_1=3$, $\beta_2=2$, $\delta_{11}=\delta_{22}=1.5$ and $\delta_{12}=\delta_{21}=0.75$

Acknowledgements

We gratefully acknowledge financial support from the Agence Nationale de Recherche ANR-05-BLAN-0024-01. This research has also received financial support from the European Research Council under the European Community Seventh Framework Program FP7/2007-2013 grant agreement N_295298. We would like to thank the editor, the associate editor and two anonymous referees for helpful comments and participants at various conferences and seminars for helpful discussions.

Bibliography

Ackerberg, D., Chen, X., Hahn, J., and Liao, Z. (2014). Asymptotic efficiency of semiparametric two-step gmm. *Review of Economic Studies*, 81(3):919–943.

Ai, C. and Chen, X. (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843.

Andrews, D. W., Berry, S., and Jia, P. (2004). Confidence regions for parameters in discrete games with multiple equilibria, with an application to discount chain store location. *WP Cowles*.

Andrews, D. W. and Shi, X. (2013). Inference based on conditional moment inequalities. *Econometrica*, 81(2):609–666.

Andrews, D. W. and Soares, G. (2010). Inference for parameters defined by moment inequalities using generalized moment selection. *Econometrica*, 78(1):119–157.

Andrews, D. W. K. and Barwick, P. J. (2012). Inference for parameters defined by moment inequalities: A recommended moment selection procedure. *Econometrica*, 80(6):2805–2826.

Andrews, I., Roth, J., and Pakes, A. (Forthcoming). Inference for linear conditional moment inequalities. *Review of Economic Studies*.

Aradillas-Lopez, A. and Rosen, A. M. (2022). Inference in ordered response games with complete information. *Journal of Econometrics*, 226(2):451–476.

Armstrong, T. B. (2014). Weighted ks statistics for inference on conditional moment inequalities. *Journal of Econometrics*, 181(2):92–116.

Armstrong, T. B. and Chan, H. P. (2016). Multiscale adaptive inference on conditional moment inequalities. *Journal of Econometrics*, 194(1):24–43.

Bajari, P., Hong, H., Krainer, J., and Nekipelov, D. (2006). Estimating static models of strategic interaction. Technical report, National Bureau of Economic Research.

Bajari, P., Hong, H., and Ryan, S. P. (2010). Identification and estimation of a discrete game of complete information. *Econometrica*, 78(5):1529–1568.

Beresteanu, A., Molchanov, I., and Molinari, F. (2011). Sharp identification regions in models with convex moment predictions. *Econometrica*, 79(6):1785–1821.

Beresteanu, A., Molchanvb, I., and Molinari, F. (2012). Partial identification using random set theory. *Journal of Econometrics*, 166:17–32.

Beresteanu, A. and Molinari, F. (2008). Asymptotic properties for a class of partially identified models. *Econometrica*, 76(4):763–814.

Berry, S. T. (1992). Estimation of a model of entry in the airline industry. *Econometrica: Journal of the Econometric Society*, pages 889–917.

Berry, S. T. and Compiani, G. (2022). An Instrumental Variable Approach to Dynamic Models. *The Review of Economic Studies*. rdac061.

Bjorn, P. A. and Vuong, Q. H. (1984). Simultaneous equations models for dummy endogenous variables: a game theoretic formulation with an application to labor force participation.

Bontemps, C. and Kumar, R. (2020). A geometric approach to inference in set-identified entry games. *Journal of Econometrics*, 218(2):373–389.

Bontemps, C., Magnac, T., and Maurin, E. (2012). Set identified linear models. *Econometrica*, 80(3):1129–1155.

Bresnahan, T. F. and Reiss, P. C. (1991a). Empirical models of discrete games. *Journal of Econometrics*, 48(1-2):57–81.

Bresnahan, T. F. and Reiss, P. C. (1991b). Entry and competition in concentrated markets. *Journal of political economy*, 99(5):977–1009.

Breusch, T., Qian, H., Schmidt, P., and Wyhowski, D. (1999). Redundancy of moment conditions. *Journal of Econometrics*, 91(1):89–111.

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

Chen, X., Christensen, T. M., and Tamer, E. (2018). Monte carlo confidence sets for identified sets. *Econometrica*, 86(6):1965–2018.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018a). Double/debiased machine learning for treatment and structural parameters.

Chernozhukov, V., Chetverikov, D., and Kato, K. (2018b). Inference on causal and structural parameters using many moment inequalities. *The Review of Economic Studies*, 86(5):1867–1900.

Chernozhukov, V., Hong, H., and Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models 1. *Econometrica*, 75(5):1243–1284.

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

Chesher, A. and Rosen, A. M. (2019). Structural modeling of simultaneous discrete choice. *Cemmap WP*.

Cho, J. and Russell, T. M. (2018). Simple inference on functionals of set-identified parameters defined by linear moments.

Ciliberto, F., Murry, C., and Tamer, E. (2021). Market structure and competition in airline markets. *Journal of Political Economy*, 129(11):2995–3038.

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

Cleeren, K., Verboven, F., Dekimpe, M. G., and Gielens, K. (2010). Intra-and interformat competition among discounters and supermarkets. *Marketing science*, 29(3):456–473.

Cox, G. and Shi, X. (2022). Simple Adaptive Size-Exact Testing for Full-Vector and Subvector Inference in Moment Inequality Models. *The Review of Economic Studies*. rdac015.

Eizenberg, A. (2014). Upstream innovation and product variety in the us home pc market. *Review of Economic Studies*, 81(3):1003–1045.

Gafarov, B. (2019). Inference in high-dimensional set-identified affine models.

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

Grieco, P. L. (2014). Discrete games with flexible information structures: An application to local grocery markets. *The RAND Journal of Economics*, 45(2):303–340.

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

Kitamura, Y. and Stoye, J. (2018). Nonparametric analysis of random utility models. *Econometrica*, 86(6):1883–1909.

Liu, Y. and Xie, J. (2020). Cauchy combination test: A powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529):393–402. PMID: 33012899.

Luo, Y. and Wang, H. (2017). Core determining class and inequality selection. *American Economic Review*, 107(5):274–77.

Magnolfi, L. and Roncoroni, C. (2022). Estimation of Discrete Games with Weak Assumptions on Information. *The Review of Economic Studies*. rdac058.

Manski, C. F. (1995). *Identification problems in the social sciences*. Harvard University Press.

Mazzeo, M. J. (2002). Product choice and oligopoly market structure. *RAND Journal of Economics*, pages 221–242.

Molchanov, I. and Molinari, F. (2014). Applications of random set theory in econometrics. *Annu. Rev. Econ.*, 6(1):229–251.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica: Journal of the Econometric Society*, pages 1349–1382.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245.

Rockafellar, R. (1997). *Convex Analysis*. Convex Analysis. Princeton University Press.

Romano, J. P., Shaikh, A. M., and Wolf, M. (2014). A practical two-step method for testing moment inequalities. *Econometrica*, 82(5):1979–2002.

Rosen, A. M. (2008). Confidence sets for partially identified parameters that satisfy a finite number of moment inequalities. *Journal of Econometrics*, 146(1):107–117.

Silvapulle, M. and Sen, P. (2011). *Constrained Statistical Inference: Order, Inequality, and Shape Constraints*. Wiley Series in Probability and Statistics. Wiley.

Toivanen, O. and Waterson, M. (2005). Market structure and entry: Where's the beef? *RAND Journal of Economics*, pages 680–699.

A Proof of Propositions

Proof of proposition 2

We want to prove the following statement: if A is not connected in $\Gamma(\theta)$, then A generates a redundant moment.

To show this, we are going to use the equivalent definition of connectedness. A subset $A \subset \mathcal{Y}$ is connected in $\Gamma(\theta)$ if and only if for every partition in 2 subsets A_1 and A_2 of A, there exists at least one element $y_1 \in A_1$ and $y_2 \in A_2$ that have overlapping equilibrium regions $\mathcal{R}_\omega(X, y_1)$ and $\mathcal{R}_\omega(X, y_2)$.

Assume that A is not connected in $\Gamma(\theta)$, then there exists A_1 and A_2 such that $A = A_1 \cup A_2$ and A_1 and A_2 are such that $\mathcal{R}_\omega(X, A_1) \cap \mathcal{R}_\omega(X, A_2) = \emptyset$. First, let us consider the moment inequalities generated by A_1 and A_2 separately. We have

$$P_0(Y \in A_1|X) \leq P_\eta(\varepsilon \in \mathcal{R}_\omega(X, A_1)) \text{ and } P_0(Y \in A_2|X) \leq P_\eta(\varepsilon \in \mathcal{R}_\omega(X, A_2)).$$

By combining these 2 inequalities, we have:

$$
\begin{aligned}
& P_0(Y \in A_1|X) + P_0(Y \in A_2|X) \leq P_\eta(\varepsilon \in \mathcal{R}_\omega(X, A_1)) + P_\eta(\varepsilon \in \mathcal{R}_\omega(X, A_2)) \\
\iff & P_0(Y \in A_1 \cup A_2|X) \leq P_\eta(\varepsilon \in \mathcal{R}_\omega(X, A_1)) + P_\eta(\varepsilon \in \mathcal{R}_\omega(X, A_2)) - P_\eta(\varepsilon \in \mathcal{R}_\omega(X, A_1) \cap \mathcal{R}_\omega(X, A_2)) \\
\iff & P_0(Y \in A|X) = P_0(Y \in A_1 \cup A_2|X) \leq \mathcal{R}_\omega(X, A_1) \cup \mathcal{R}_\omega(X, A_2)) = P_\eta(\varepsilon \in \mathcal{R}_\omega(X, A).
\end{aligned}
$$

The second line in the expression above comes from the fact that A_1 and A_2 are disjoint and thus $P_0(Y \in A_1 \cup A_2) = P_0(Y \in A_1) + P_0(Y \in A_2)$. What's more, by assumption $\mathcal{R}_\omega(A_1) \cap \mathcal{R}_\omega(A_2) = \emptyset$ and as a consequence, $P_\eta(\varepsilon \in \mathcal{R}_\omega(A_1) \cap \mathcal{R}_\omega(A_2)) = 0$. The last line stems from the inclusion-exclusion formula. This proves our result. ■

Proof of proposition 3

From proposition 1, we know that an outcome $y_1 = (N_1, ..., N_T)$ is a NE of this game if and only if: $\forall t \in \mathcal{T}$,

$$-\pi_t(X, N_t, \mathbf{N}_{-t}; \omega) \leq \varepsilon_t \leq -\pi_t(X, N_t + 1, \mathbf{N}_{-t}; \omega).$$

This system of $2T$ inequalities defines a region $\mathcal{R}_\omega(X, y_1)$ in the space of unobserved heterogeneity. Analogously, we can define a region $\mathcal{R}_\omega(X, y_2)$ in the space of unobserved heterogeneity where $y_2 = (\bar{N}_1, ..., \bar{N}_\tau)$ is a NE of this game. $\mathcal{R}_\omega(X, y_1)$ and $\mathcal{R}_\omega(X, y_2)$ are T-cubes in $\mathbb{R}^T$. Hence, these regions have a non-empty intersection if and only if for each dimension (which here corresponds to a type), the projections of these cubes have a non-empty intersections. Formally,

$$\mathcal{R}_\omega(X, y_1) \cap \mathcal{R}_\omega(X, y_2) \neq \emptyset \iff \forall t \ \in \mathcal{T}, \mathrm{Proj}(\mathcal{R}_\omega(X, y_1)|e_t) \cap \mathrm{Proj}(\mathcal{R}_\omega(X, y_2)|e_t) \neq \emptyset$$

Once again, by definition of $\mathcal{R}_\omega(X, y_1)$, $\mathrm{Proj}(\mathcal{R}_\omega(X, y_1)|e_t) = [-\pi_t(X, N_t, \mathbf{N}_{-t}; \omega) \leq \varepsilon_t \leq -\pi_t(X, N_t + 1, \mathbf{N}_{-t}; \omega)[$. Likewise, $\mathrm{Proj}(\mathcal{R}_\omega(X, y_2)|e_t) = [-\pi_t(X, \bar{N}_t, \bar{\mathbf{N}}_{-t}; \omega) \leq \varepsilon_t \leq -\pi_t(X, \bar{N}_t + 1, \bar{\mathbf{N}}_{-t}; \omega)[$

Furthermore, from basic analysis results on sets in $\mathbb{R}$, $[a, b[\cap [c, d[\neq \emptyset \iff a < d$ and $c < b$, which applied to our two regions, proves the result. ■

Proof of proposition 5

(i) Following Equation (8), if $\theta \in \Theta_I$, for any $j = 1, \ldots, p$,

$$\mathbb{E} m_j(Y, \theta) \geq 0.$$

Therefore, $\exp(-\rho(C_{\theta,j} - q_j^\top P_0(Y))) \leq 1$ because ρ is positive. As a result

$$\log\left(\frac{1+\sum_{j=1}^{p}\exp(-\rho(C_{\theta,j} - q_j^\top P_0(Y)))}{p+1}\right) \leq 0,$$

and $G_{\theta,\rho}(P_0(Y)) = -\rho^{-1}\log\left(\frac{1+\sum_{j=1}^{p}\exp(-\rho(C_{\theta,j} - q_j^\top P_0(Y)))}{p+1}\right) \geq 0$. Which proves $\theta \in \Theta_I^o(\rho)$.

(ii) Likewise for (i), we prove similarly $A(\theta) \subset A_\rho^o(\theta)$. $P \mapsto -G_{\theta,\rho}(P)$ is a strictly convex function. If P_1 and P_2 belong to $A_\rho^o(\theta)$, take $\lambda \in (0,1)$.

$$\begin{aligned}-G_{\theta,\rho}(\lambda P_1 + (1-\lambda)P_2) &< -\lambda G_{\theta,\rho}(P_1) - (1-\lambda)G_{\theta,\rho}(P_2)\\ &< 0 \iff (\lambda P_1 + (1-\lambda)P_2) \in A_\rho^o(\theta).\end{aligned}$$

The strict convexity of $A_\rho^o(\theta)$ comes from the strict inequality in the above inequality. Therefore, if both P_1 and P_2 belongs to the frontier of $A_\rho^o(\theta)$, i.e., $G_{\theta,\rho}(P_1) = G_{\theta,\rho}(P_2) = 0$, no other point of the segment $[P_1 P_2]$ lies on the frontier.

(iii) TBD

■

Proof of proposition 6

Before expanding the test-statistic $\tilde{\xi}_n(\theta)$, we need a few lemmas to characterize the higher order derivatives of

$$G_{\theta,\rho}: \ P \mapsto G_{\theta,\rho}(P).$$

Lemma 1 *The function $G_{\theta,\rho}: \ P \mapsto G_{\theta,\rho}(P)$ is infinitely differentiable in any $P \in \mathbb{R}^T$. Furthermore,*

- *we have a close form expression for the gradient:*

$$\nabla G_{\theta,\rho}(P_0) = -\sum_{j=1}^{p} \mu_j q_j, \tag{A.1}$$

in which

$$\mu_j = \frac{\exp(-\rho(C_{\theta,j} - q_j^\top P_0))}{1+\sum_{j=1}^{p}\exp(-\rho(C_{\theta,j} - q_j^\top P_0))}.$$

- *as well as for the Hessian:*

$$H_{G_{\theta,\rho}}(P_0) = \nabla G_{\theta,\rho}(P_0)\nabla^\top = \rho \left(\sum_{j=1}^{p} \mu_j q_j q_j^\top + \left(\sum_{j=1}^{p} \mu_j q_j \right) \left(\sum_{j=1}^{p} \mu_j q_j^\top \right) \right). \tag{A.2}$$

The proof is straightforward and thus omitted. Observe that the Hessian in P_0 is equal to ρ times a bounded matrix.

Lemma 2 *Let*

$$\mathcal{J}(\theta) = \{j \in \{1, \dots, p\} \,|\mathbb{E} m_j(Y, \theta) = 0\}$$

and $J_0 = card(\mathcal{J}(\theta))$*. Then,*

$$\lim_{\rho\to\infty} \nabla G_{\theta,\rho}(P_0) = \Gamma_0(\theta) = \frac{-\sum_{j\in\mathcal{J}(\theta)} q_j}{1 + J_0}$$

and

$$\|\nabla G_{\theta,\rho}(P_0) - \Gamma_0(\theta)\|_2 \xrightarrow[\rho\to\infty]{P} 0.$$

Proof

Obviously, for any positive value of ρ;

$$\exp(-\rho(C_{\theta,j} - q_j^\top P)) = 1 \text{ if } \mathbb{E} m_j(Y, \theta) = 0.$$

The first part follows.

Moreover,

$$\|\nabla G_{\theta,\rho_n}(P_n) - \Gamma_0(\theta)\|_2 \le \|\nabla G_{\theta,\rho_n}(P_n) - \nabla G_{\theta,\rho_n}(P_0)\|_2 + \|\nabla G_{\theta,\rho_n}(P_0) - \Gamma_0(\theta)\|_2\,, \tag{A.3}$$

$$= \left\| H_{G_{\theta,\rho}}(\tilde{P}_n)(P_n - P_0) \right\|_2 + \|\nabla G_{\theta,\rho_n}(P_0) - \Gamma_0(\theta)\|_2\,, \tag{A.4}$$

in which $\tilde{P}$ is a point in the segment $[P_n P_0]$.

Following, Lemma 1, the first term is equal to:

$$H_{G_{\theta,\rho}}(\tilde{P})(P_n - P_0) = \frac{\rho_n}{\sqrt{n}} \left(\sum_{j=1}^{p} \tilde{\mu}_j q_j q_j^\top Z_n + \left(\sum_{j=1}^{p} \tilde{\mu}_j q_j \right) \left(\sum_{j=1}^{p} \tilde{\mu}_j q_j^\top Z_n \right) \right)$$

in which $Z_n = \sqrt{n}(P_n - P_0)$ and

$$\tilde{\mu}_j = \frac{\exp(-\rho(C_{\theta,j} - q_j^\top \tilde{P}_n))}{1 + \sum_{j=1}^p \exp(-\rho(C_{\theta,j} - q_j^\top \tilde{P}_n))}.$$

Therefore

$$\left\| H_{G_{\theta,\rho}}(\tilde{P}_n)(P_n - P_0) \right\|_2 \leq \frac{\rho_n}{\sqrt{n}} \times O_p(1),$$

and it tends to 0 in probability given the speed of divergence of ρ_n.

The second term in (A.4) tends trivially to 0 in probability from the definition of $\Gamma_0(\theta)$. ■

We now finish the proof of the proposition. First, expand $\tilde{\xi}_n(\theta)$ around P_0. In the following, $Z_n = \sqrt{n}(P_n - P_0)$. As P_n is a frequency estimator, Z_n is asymptotically normally distributed with mean 0 and variance Σ_0 (we denote its limit Z_∞).

$$\begin{aligned}
\tilde{\xi}_n(\theta) &= \sqrt{n} G_{\theta,\rho_n}(P_n) && \text{(A.5)}\\
&= \sqrt{n} G_{\theta,\rho_n}(P_0) + \nabla G_{\theta,\rho_n}(P_0) Z_n + \frac{1}{\sqrt{n}} Z_n^\top H_{G_{\theta,\rho_n}}(\tilde{P}_n) Z_n && \text{(A.6)}\\
&= \sqrt{n} G_{\theta,\rho_n}(P_0) + \Gamma_0(\theta) Z_n + \left(\nabla G_{\theta,\rho_n}(P_0) - \Gamma_0(\theta)\right) Z_n + \frac{\rho_n}{\sqrt{n}} O_p(1). && \text{(A.7)}\\
&= \sqrt{n} G_{\theta,\rho_n}(P_0) + \Gamma_0(\theta) Z + \Gamma_0(\theta)(Z_n - Z) + \left(\nabla G_{\theta,\rho_n}(P_0) - \Gamma_0(\theta)\right) Z_n + O_p(\frac{\rho_n}{\sqrt{n}}). && \text{(A.8)}
\end{aligned}$$

The third term in (A.8) is $O_P(1/\sqrt{n})$ by the CLT, the fourth term is $o_P(1)$ because product of a term tending to 0 with a bounded one (in probability), the last term tends to 0 giventhe speed of convergence of ρ_n. The second term is the normal with variance $\Gamma_0(\theta)\Sigma_0\Gamma_0(\theta)^\top$. Hence, we get the result. ■

Proof of proposition 7

- First, we show it with p^* instead of $\widehat{p^*}_n$ in Equation(14). Remember that Proposition 6 allows us to have the following expansion:

$$\sqrt{n}(G_{\theta,\rho_n}(P_n) - G_{\theta,\rho_n}(P_0) = \Gamma_0(\theta) Z + o_P(1),$$

under the restrictions on ρ_n. Furthermore,

$$
\begin{aligned}
G_{\theta,\rho_n}(P_0) &= \frac{-1}{\rho_n}\log\left(\frac{1+\sum_{j=1}^{p}\exp(-\rho_n(C_{\theta,j}-q_j^\top P_0))}{p+1}\right)\\
&= \frac{-1}{\rho_n}\log\left(\frac{1+p^*}{1+p}+\frac{\sum_{j\notin\mathcal{J}(\theta)}\exp(-\rho_n(C_{\theta,j}-q_j^\top P_0))}{p+1}\right)\\
&= \frac{-1}{\rho_n}\log\left(\frac{1+p^*}{1+p}\right)-\frac{1}{\rho_n}\log\left(1+\frac{\sum_{j\notin\mathcal{J}(\theta)}\exp(-\rho_n(C_{\theta,j}-q_j^\top P_0))}{1+p^*}\right)
\end{aligned}
$$

We need to show that the second term tends to zero at any polynomial rate. Take $\eta = \min_{j\notin\mathcal{J}(\theta)} C_{\theta,j} - q_j^\top P_0$. By definition of $\mathcal{J}(\theta)$, $\eta > 0$. Therefore,

$$
\sum_{j\notin\mathcal{J}(\theta)}\exp(-\rho_n(C_{\theta,j}-q_j^\top P_0)) \leq p\exp(-\rho_n\eta).
$$

The last term tends to 0 when n tends to infinity.

and, using $x - x^2/2 \leq \log(1+x) \leq x$ in a neighborhood of 0, we obtain:

$$
\frac{W_n - W_n^2/2}{\rho_n} \leq \frac{1}{\rho_n}\log\left(1+\frac{\sum_{j\notin\mathcal{J}(\theta)}\exp(-\rho_n(C_{\theta,j}-q_j^\top P_0))}{1+p^*}\right) \leq \frac{W_n}{\rho_n},
$$

with $W_n = \frac{p}{p+1}\exp(-\rho_n\eta)$.

Therefore, for any $b > 0$,

$$
\frac{1}{\rho_n}\log\left(1+\frac{\sum_{j\notin\mathcal{J}(\theta)}\exp(-\rho_n(C_{\theta,j}-q_j^\top P_0))}{1+p^*}\right) = o(1/n^b),
$$

Then we have under H_0:

$$
\sqrt{n}\left(G_{\theta,\rho_n}(P_n) + \frac{1}{\rho_n}\log\left(\frac{1+p^*}{1+p}\right)\right) = \Gamma_0(\theta)Z + o_P(1),
$$

which achieves the first part of the proof.

- Next, we show that the estimation of p^* does not change the expansion above. Remember the definition of $\widehat{p^*}_n$

$$
\widehat{p^*}_n = \sum_{j=1}^{p}\mathbf{1}\left\{\frac{C_{\theta,j}-q_j^\top P_n}{\sqrt{q_j^\top \Sigma_n q_j}} < \tau_n\right\}.
$$

Using the expansion above, we can rewrite it with $\widehat{p^*}_n$:

$$\sqrt{n}\left(G_{\theta,\rho_n}(P_n) + \frac{1}{\rho_n}\log\left(\frac{1+\widehat{p^*}_n}{1+p}\right)\right) + \frac{\sqrt{n}}{\rho_n}\log\left(\frac{1+p^*}{1+\widehat{p^*}_n}\right) = \Gamma_0(\theta)Z + o_P(1).$$

We need to show that the second term of the left hand side of the equality above tends to 0.

Observe that

$$P_0(\frac{\sqrt{n}}{\rho_n}\log\left(\frac{1+p^*}{1+\widehat{p^*}_n}\right) \neq 0) = P_0(\widehat{p^*}_n \neq p^*).$$

Then, we calculate the probability for a binding moment to not be selected:

$$\begin{aligned}
P_{ns} &= P((Z/\sqrt{n} + o_P(1/\sqrt{n})) > \tau_n) \\
&= 1 - \Phi(\sqrt{n}\tau_n + o_P(1)) \\
&\leq K\frac{1}{\sqrt{n}\tau_n}\exp(-n\tau_n^2/2) \text{following classical result on the upper tail of the } N(0,1)
\end{aligned}$$

Similarly the probability for a non-binding moment (its expectation is denoted $\mu > 0$) to be selected is equal to

$$\begin{aligned}
P_s &= P((Z/\sqrt{n} + \mu + o_P(1/\sqrt{n})) < \tau_n) \\
&= \Phi(\sqrt{n}(\tau_n - \mu) + o_P(1))
\end{aligned}$$

τ_n tending toward 0 and μ being positive, the quantity inside $\Phi(\cdot)$ tends to $-\infty$. The probability tends to 0 at any polynomial rate.

As a result, for any $b > 0$,

$$n^b P_0(\widehat{p^*}_n \neq p^*) \xrightarrow[n\to\infty]{P} 0.$$

■

Proof of proposition 8

- The asymptotic validity is a direct corollary of Proposition 7.
  - Consistency

If $\theta \notin \Theta_I$,

$$\min_{j=1,\ldots,p} \mathbb{E} m_j(Y,\theta) = \gamma < 0.$$

Assume without loss of generality, that this is the first moment, and, for simplicity that this is the only violation. The proof can easily adapt to the general case.

Then,

$$plim\left(C_{\theta,1} - q_1^\top P_n < \gamma/2\right) = 1.$$

Or, equivalently,

$$Pr\left(\frac{1+\sum_{j=1}^{p}\exp(-\rho_n(C_{\theta,j} - q_j^\top P))}{p+1} > \frac{1+\exp(-\rho_n\gamma/2)}{p+1}\right) \xrightarrow[n\to\infty]{P} 1,$$

i.e.,

$$Pr\left(G_{\theta,\rho_n}(P_n) < -\frac{1}{\rho_n}\log\left(\frac{1+\exp(-\rho_n\gamma/2)}{p+1}\right)\right) \xrightarrow[n\to\infty]{P} 1.$$

The upper bound in the inequality above is equal to:

$$\frac{\gamma}{2} - \frac{1}{\rho_n}\log(1+\exp(\rho_n\gamma/2)) + \frac{1}{\rho_n}\log(1+p),$$

which is below a fixed negative quantity when n is sufficiently large. Thus,

$$\Pr\left(\sqrt{n}\frac{G_{\theta,\rho_n}(P_n) - \frac{1}{\rho_n}\log\left(\frac{1+p}{1+\widehat{p^*}_n}\right)}{\sqrt{V_n}} \geq z_\alpha\right) \xrightarrow[n\to\infty]{P} 0.$$

■

Proof of proposition 9

To show this, we just need to prove that the graph generated by $\Gamma(\theta, X)$ remains the same for all $X \in \mathcal{X}$. The graph $\Gamma(\theta, X)$ is such that there exists an edge between two elements y_1 and y_2 if their equilibrium regions $\mathcal{R}_\omega(X, y_1)$ and $\mathcal{R}_\omega(X, y_2)$ overlap. Now from proposition 3, we know that this is the case if

$\forall t \in \mathcal{T}$, such that $0 \leq N_t, \bar{N}_t \leq a$:

$$\begin{cases} -\pi_t(X, N_t, \mathbf{N_{-t}}; \omega) < -\pi_t(X, \bar{N}_t + 1, \bar{\mathbf{N}}_{-t}; \omega) \\ -\pi_t(X, \bar{N}_t, \bar{\mathbf{N}}_{-t}; \omega) < -\pi_t(X, N_t + 1, \mathbf{N}_{-t}; \omega). \end{cases}$$

From the decomposition in Assumption 3. The previous conditions become:

$$\begin{cases} -\kappa_t(N_t, \mathbf{N_{-t}}; \omega_2) < -\kappa_t(\bar{N}_t + 1, \bar{\mathbf{N}}_{-t}; \omega_2) \\ -\kappa_t(\bar{N}_t, \bar{\mathbf{N}}_{-t}; \omega_2) < -\kappa_t(N_t + 1, \mathbf{N}_{-t}; \omega_2). \end{cases}$$

These conditions do not depend on X. ■

Inference with covariates

The smoothed max approach

Let us consider the kernel estimator of h_0,

$$\hat{h}_0(x) = \frac{\sum_{i=1}^n K_\sigma(x - X_i)\mathbf{1}\{Y_i = y\}}{\sum_{i=1}^n K_\sigma(x - X_i)},$$

where σ is the bandwidth and $K_\sigma(x) = \frac{1}{\sigma^d} K\left(\frac{x}{\sigma}\right)$. Lets also define $\hat{f}_0(x) = \frac{1}{n}\sum_{i=1}^n K_\sigma(x - X_i)$ and $\hat{w}_0(x) = \frac{1}{n}\sum_{i=1}^n K_\sigma(x - X_i)\mathbf{1}\{Y_i = y\}$. Under assumption 5, we have

$$\sqrt{n}||\hat{h}_0 - h_0||^2 \xrightarrow{P} 0 \quad \text{and} \quad \sqrt{n}||\hat{f}_0 - f_0||^2 \xrightarrow{P} 0,$$

where $\|\cdot\|$ is a Sobolev norm, for non-negative integer k, defined as

$$\|f\| = \max_{j \leq k} \max_{x \in \mathcal{X}} \left\| \frac{\partial^j f(x)}{\partial x^j} \right\|.$$

Then, from Newey and McFadden (1994) Lemma 8.9, we have

$$\|f_0 - \mathbb{E}(\hat{f}_0)\| = O\left(\sigma^m\right),$$

if $m + k < \alpha$. Similar result holds for $\hat{h}_0$. For the variance term, following Newey and McFadden (1994), we have

$$\|\hat{f}_0 - \mathbb{E}(\hat{f}_0)\| = O_P\left[\left(\frac{\ln n}{n\sigma^{d+2k}}\right)^{\frac{1}{2}}\right].$$

Similar result holds for $\hat{w}_0$.

Lemma 3 *Under assumption 5, we have*

$$\left\|\hat{h}_0 - \mathbb{E}(\hat{h}_0)\right\| = O_P\left[\left(\frac{\ln n}{n\sigma^{d+2k}}\right)^{\frac{1}{2}}\right].$$

Lemmas

The next two lemmas verify technical conditions for kernel estimator which are needed to derive asymptotic distribution.

Lemma 4 (Mean Square Differentiability) *Under assumption 5, we have uniformly over smoothing parameter* $\rho > 0$

$$\sqrt{n}\left[\int \frac{\partial G_{\theta,\rho}(x,h_0)}{\partial h}^{\top}(\hat{h}_0(x) - h_0(x))\mathrm{d}x - \frac{1}{n}\sum_{i=1}^{n}\alpha(W_i)\right] \xrightarrow{P} 0.$$

Proof:

We have

$$\begin{aligned}
&\int \frac{\partial G_{\theta,\rho}(x,h_0)}{\partial h}^{\top}(\hat{h}_0(x) - h_0(x))\mathrm{d}x \\
&= \int \frac{\partial G_{\theta,\rho}(x,h_0)}{\partial h}^{\top}\left(\frac{\sum_{i=1}^{n}K_\sigma(x-X_i)\mathbf{1}\{Y_i=y\}}{\sum_{i=1}^{n}K_\sigma(x-X_i)} - h_0(x)\right)\mathrm{d}F_0(x) \\
&= \sum_{i=1}^{n}\int \frac{K_\sigma(x-X_i)}{\sum_{i=1}^{n}K_\sigma(x-X_i)}\underbrace{\frac{\partial G_{\theta,\rho}(x,h_0)}{\partial h}^{\top}(\mathbf{1}\{Y_i=y\} - h_0(x))}_{\alpha(w)}\mathrm{d}F_0(x) \\
&= \int \alpha(w)d\tilde{F}(w) \\
&= \frac{1}{n}\sum_{i=1}^{n}\int \frac{K_\sigma(x-X_i)}{\frac{1}{n}\sum_{i=1}^{n}K_\sigma(x-X_i)}\underbrace{\frac{\partial G_{\theta,\rho}(x,h_0)}{\partial h}^{\top}[-h_0(x)\ I]}_{\beta(x)}\underbrace{\begin{bmatrix}\mathbf{1}\\ \mathbf{1}\{Y_i=y\}\end{bmatrix}}_{Z_i}\mathrm{d}F_0(x),
\end{aligned}$$

where distribution $\tilde{F}$ is mix of empirical conditional measure and density f_0. The random variable X follows density $f_0(\cdot)$ and conditional on $X=x$, $\mathbf{1}\{Y_i=y\}$ is distributed with empirical measure with points mass $\frac{K_\sigma(x-X_i)}{\sum_{i=1}^{n}K_\sigma(x-X_i)}$. So, we just need to show that

$$\sqrt{n}\left[\int \alpha(w)d\tilde{F}(w) - \int \alpha(w)d\hat{F}(w)\right] \xrightarrow{P} 0,$$

where $\hat{F}$ is just empirical distribution. We have

$$
\begin{aligned}
&\sqrt{n}\left[\int \alpha(w) d\tilde{F}(w)-\int \alpha(w) d\hat{F}(w)\right] \\
&=\frac{1}{\sqrt{n}} \sum_{i=1}^{n}\left[\int \frac{K_\sigma(x-X_i)}{\frac{1}{n}\sum_{i=1}^{n} K_\sigma(x-X_i)} \beta(x) \mathrm{d} F_0(x)-\beta(X_i)\right] Z_i \\
&=\frac{1}{\sqrt{n}} \sum_{i=1}^{n}\left[\int \frac{K_\sigma(x-X_i)}{f_0(x)} \beta(x) f_0(x) \mathrm{d} x-\beta(X_i)\right] Z_i \\
&\quad+\frac{1}{\sqrt{n}} \sum_{i=1}^{n}\left[\int \frac{K_\sigma(x-X_i)(f_0(x)-\hat{f}_0(x))}{\hat{f}_0(x)} \beta(x) \mathrm{d} x\right] Z_i \\
&=\frac{1}{\sqrt{n}} \sum_{i=1}^{n}\left[\int \frac{K_\sigma(x-X_i)}{f_0(x)} \beta(x) f_0(x) \mathrm{d} x-\beta(X_i)\right] Z_i \\
&\quad+\sqrt{n}\left[\int (f_0(x)-\hat{f}_0(x)) \beta(x)\left\{\frac{\frac{1}{n}\sum_{i=1}^{n} K_\sigma(x-X_i) Z_i}{\hat{f}_0(x)}-\left[\begin{array}{c}\mathbf{1} \\ h_0(x)\end{array}\right]\right\} \mathrm{d} x\right] \\
&\quad+\sqrt{n}\underbrace{\left[\int (f_0(x)-\hat{f}_0(x)) \beta(x)\left[\begin{array}{c}\mathbf{1} \\ h_0(x)\end{array}\right] \mathrm{d} x\right]}_{=0} \\
&=\frac{1}{\sqrt{n}} \sum_{i=1}^{n}\left[\int \frac{K_\sigma(x-X_i)}{f_0(x)} \beta(x) f_0(x) \mathrm{d} x-\beta(X_i)\right] Z_i \\
&\quad+\sqrt{n} \int \frac{\partial G_{\theta, \rho}(x, h_0)}{\partial h}^{\top}(f_0(x)-\hat{f}_0(x))(h_0(x)-\hat{h}_0(x)) \mathrm{d} x .
\end{aligned}
$$

For the first term, we will use Chebyshev's inequality and show its expectation and variance converge to zero. Let $\gamma_0(x)=\mathbb{E}(Z_i|X=x)$, then

$$
\begin{aligned}
&\left\|\sqrt{n}\mathbb{E}\left[\left\{\int K_\sigma(x-X_i)\beta(x)\mathrm{d}x-\beta(X_i)\right\} Z_i\right]\right\| \\
&\quad=\sqrt{n}\left\|\int\left\{\int \beta(x+\sigma u) K(u) \mathrm{d} u\right\} \gamma_0(x) f_0(x) \mathrm{d} x-\int \beta(x) \gamma_0(x) f_0(x) \mathrm{d} x\right\| \\
&\quad=\sqrt{n}\left\|\iint \beta(x) K(u) \gamma_0(x-\sigma u) f_0(x-\sigma u) \mathrm{d} u \ \mathrm{d} x-\int \beta(x) \gamma_0(x) f_0(x) \mathrm{d} x\right\| \\
&\quad=\sqrt{n}\left\|\int \beta(x)\left\{\int\left[f_0(x-\sigma u) \gamma_0(x-\sigma u)-f_0(x) \gamma_0(x)\right] K(u) \mathrm{d} u\right\} \mathrm{d} x\right\| \\
&\quad\leqslant \sqrt{n} \int\|\beta(x)\|\left\|\int\left[f_0(x-\sigma u) \gamma_0(x-\sigma u)-f_0(x) \gamma_0(x)\right] K(u) \mathrm{d} u\right\| \mathrm{d} x \leqslant C \sqrt{n} \sigma^m \int\|\beta(x)\| \mathrm{d} x,
\end{aligned}
$$

where last bound, for some constant C, follows from Taylor expansion of $f_0(x-\sigma u)$, $\gamma_0(x-\sigma u)$ and assumption (5)(1). Therefore, $\left\|\sqrt{n}\mathbb{E}\left[\left\{\int \beta(x)K_\sigma(x-X_i)\mathrm{d}x-\beta(X_i)\right\} Z_i\right]\right\| \leqslant C\sqrt{n}\sigma^m \to 0$.

Also, by almost everywhere continuity of $\beta(x), \beta(x+\sigma u) \rightarrow \beta(x)$ for almost all x. Also, on the bounded support of $K(u)$, for small enough $\sigma, \beta(x+\sigma u) \leqslant \sup_{\|v\| \leqslant \varepsilon} \beta(x+v)$, so by the dominated convergence theorem, $\int \beta(x+\sigma u) K(u) \mathrm{d} u \rightarrow \int \beta(x) K(u) \mathrm{d} u=\beta(x)$ for almost all x. The boundedness of $K(u)$ and dominated convergence theorem gives

$$
\mathbb{E}\left[\left\|\int \beta(x) K_\sigma\left(x-X_i\right) \mathrm{d} x-\beta\left(X_i\right)\right\|^4\right] \rightarrow 0,
$$

so by the CauchySchwartz inequality, $\mathbb{E}\left[\left\|Z_i\right\|^2\left\|\int \beta(x) K_\sigma\left(x-X_i\right) \mathrm{d} x-\beta\left(X_i\right)\right\|^2\right] \rightarrow 0$. Mean square differentiability condition follows from the Chebyshev inequality, since the mean and variance of $n^{-1/2} \sum_{i=1}^n\left[\int \beta(x) K_\sigma\left(x-X_i\right) \mathrm{d} x-\beta\left(X_i\right)\right] Z_i$ go to zero.

The second term in the expression is bounded, uniformly over ρ, by

$$
\begin{aligned}
&\left|\sqrt{n} \int \frac{\partial G_{\theta, \rho}(x, h_0)}{\partial h}^{\top}(f_0(x)-\hat{f}_0(x))(h_0(x)-\hat{h}_0(x)) \mathrm{d} x\right| \\
&\quad \leq \sqrt{n} \int\left|\frac{\partial G_{\theta, \rho}(x, h_0)}{\partial h}^{\top}(f_0(x)-\hat{f}_0(x))(h_0(x)-\hat{h}_0(x))\right| \mathrm{d} x \\
&\quad \leq \sqrt{n}\|f_0(x)-\hat{f}_0(x)\| \int\left|\frac{\partial G_{\theta, \rho}(x, h_0)}{\partial h}^{\top}(h_0(x)-\hat{h}_0(x))\right| \mathrm{d} x \\
&\quad \leq \sqrt{n}\|f_0(x)-\hat{f}_0(x)\| \int\left|\frac{\partial G_{\theta, \rho}(x, h_0)}{\partial h}^{\top}(h_0(x)-\hat{h}_0(x))\right| \mathrm{d} x \\
&\quad \leq \sqrt{n}\|\hat{h}_0-h_0\|\|\hat{f}_0-f_0\| \int \sum_i \frac{\partial G_{\theta, \rho}(x, h_0)}{\partial h_i} d x \\
&\quad =|\mathcal{X}| \sqrt{n}\|\hat{h}_0-h_0\|\|\hat{f}_0-f_0\| .
\end{aligned}
$$

$\sqrt{n}\|\hat{h}_0-h_0\|\|\hat{f}_0-f_0\|$, which converges to zero in probability. ■

Lemma 5 (Stochastic Equicontinuity) *Under assumption 5, we have uniformly over smoothing parameter $\rho>0$*

$$
\sqrt{n}\left[\frac{1}{n} \sum_{i=1}^n \frac{\partial G_{\theta, \rho}(X_i, h_0)}{\partial h}^{\top}(\hat{h}_0(X_i)-h_0(X_i))-\int \frac{\partial G_{\theta, \rho}(x, h_0)}{\partial h}^{\top}(\hat{h}_0(x)-h_0(x)) \mathrm{d} F_0(x)\right] \xrightarrow{P} 0 .
$$

Proof: It is easy to see that

$$\begin{aligned}
\hat{h}_0(x) - h_0(x) &= \frac{\hat{w}_0(x)}{\hat{f}_0(x)} - h_0(x) \\
&= \frac{\hat{w}_0(x) - h_0(x)\hat{f}_0(x)}{\hat{f}_0(x)} \\
&= \frac{\hat{w}_0(x) - h_0(x)\hat{f}_0(x)}{\hat{f}_0(x)} \left[\frac{\hat{f}_0(x)}{f_0(x)} + 1 - \frac{\hat{f}_0(x)}{f_0(x)}\right] \\
&= \underbrace{\frac{\hat{w}_0(x) - h_0(x)\hat{f}_0(x)}{f_0(x)}}_{I} + \underbrace{\frac{(\hat{h}_0(x) - h_0(x))(\hat{f}_0(x) - f_0(x))}{f_0(x)}}_{II}.
\end{aligned}$$

Based on this, we divide the problem into two sub-problems I and II. For the sub-problem I, we need to show[24]

$$\begin{aligned}
&\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} \frac{\frac{\partial G_{\theta,\rho}(X_i,h_0)}{\partial h}^\top}{f_0(X_i)}(\hat{w}_0(X_i) - h_0(X_i)\hat{f}_0(X_i)) - \int \frac{\partial G_{\theta,\rho}(x,h_0)}{\partial h}^\top (\hat{w}_0(x) - h_0(x)\hat{f}_0(x))\mathrm{d}x\right] \\
&= \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n} \frac{\frac{\partial G_{\theta,\rho}(X_i,h_0)}{\partial h}^\top}{f_0(X_i)}[-h_0(X_i)\ I]\begin{bmatrix}\hat{f}_0(X_i)\\ \hat{w}_0(X_i)\end{bmatrix} - \int \frac{\partial G_{\theta,\rho}(x,h_0)}{\partial h}^\top [-h_0(x)\ I]\begin{bmatrix}\hat{f}_0(x)\\ \hat{w}_0(x)\end{bmatrix}\mathrm{d}x\right] \\
&= \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\beta(X_i)}{f_0(X_i)}\frac{1}{n}\sum_{j=1}^{n}K_\sigma(X_i - X_j)Z_j - \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}_X\left[\frac{\beta(X)}{f_0(X)}K_\sigma(X - X_j)Z_j\right]\right] \\
&= \sqrt{n}\Bigg[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\beta(X_i)}{f_0(X_i)}K_\sigma(X_i - X_j)Z_j - \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}_X\left[\frac{\beta(X)}{f_0(X)}K_\sigma(X - X_j)Z_j\right] \\
&\quad - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{X,Z}\left[\frac{\beta(X_i)}{f_0(X_i)}K_\sigma(X_i - X)Z\right] + \mathbb{E}\left[\frac{\beta(X_i)}{f_0(X_i)}K_\sigma(X_i - X_j)Z_j\right]\Bigg] \\
&\quad + \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{X,Z}\left[\frac{\beta(X_i)}{f_0(X_i)}K_\sigma(X_i - X)Z\right] - \mathbb{E}\left[\frac{\beta(X_i)}{f_0(X_i)}K_\sigma(X_i - X_j)Z_j\right]\right]
\end{aligned}$$

We will apply Lemma 6 from Newey and McFadden (1994) to verify that first bracket converges to

[24] We used the notation $\beta(x) = \frac{\partial G_{\theta,\rho}(x,h_0)}{\partial h}^\top [-h_0(x)\ I]$ and $Z_i = \begin{bmatrix}\mathbf{1}\\ \mathbf{1}\{Y_i = y\}\end{bmatrix}$

0 in probability, uniformly over ρ. It is easy to see that uniformly over ρ, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{\beta(X_i)}{f_0(X_i)}K_\sigma(X_i - X_i)Z_i\right\|\right] &= \mathbb{E}\left[\left\|\frac{\frac{\partial G_{\theta,\rho}(X_i,h_0)}{\partial h}^\top}{f_0(X_i)}[-h_0(X_i)\ I]K_\sigma(0)Z_i\right\|\right] \\
&= K_\sigma(0)\mathbb{E}\left[\left\|\frac{\frac{\partial G_{\theta,\rho}(X_i,h_0)}{\partial h}^\top}{f_0(X_i)}[-h_0(X_i)\ I]\begin{bmatrix}\mathbf{1}\\ \mathbf{1}\{Y_i = y\}\end{bmatrix}\right\|\right] \\
&\leq K_\sigma(0)\mathbb{E}\left[\left\|\frac{\frac{\partial G_{\theta,\rho}(X_i,h_0)}{\partial h}^\top}{f_0(X_i)}\begin{bmatrix}\mathbf{1}\\ \mathbf{1}\end{bmatrix}\right\|\right] \\
&\leq CK_\sigma(0),
\end{aligned}$$

where second last bound uses the fact that $-1 \leq \mathbf{1}\{Y_i = y_k\} - h_{0,k}(X_i) \leq 1$ and $\frac{\partial G_{\theta,\rho}(X_i,h_0)}{\partial h} \geq 0$. Similarly, uniformly over ρ, we have

$$\begin{aligned}
\mathbb{E}\left[\left\|\frac{\beta(X_i)}{f_0(X_i)}K_\sigma(X_i - X_j)Z_j\right\|^2\right] &= \mathbb{E}\left[\left\|\frac{\frac{\partial G_{\theta,\rho}(X_i,h_0)}{\partial h}^\top}{f_0(X_i)}[-h_0(X_i)\ I]K_\sigma(X_i - X_j)Z_j\right\|^2\right] \\
&\leq M\mathbb{E}\left[\left\|\frac{\frac{\partial G_{\theta,\rho}(X_i,h_0)}{\partial h}^\top}{f_0(X_i)}[-h_0(X_i)\ I]\begin{bmatrix}\mathbf{1}\\ \mathbf{1}\{Y_j = y\}\end{bmatrix}\right\|^2\right] \\
&\leq M\mathbb{E}\left[\left\|\frac{\frac{\partial G_{\theta,\rho}(X_i,h_0)}{\partial h}^\top}{f_0(X_i)}\begin{bmatrix}\mathbf{1}\\ \mathbf{1}\end{bmatrix}\right\|\right] \\
&\leq CM,
\end{aligned}$$

where M is bound on the kernel.

Second big bracket will converge in probability to zero if

$$\mathbb{E}\left[\left\|\mathbb{E}_{X,Z}\left[\frac{\beta(X_i)}{f_0(X_i)}K_\sigma(X_i - X)Z\right]\right\|^2\right] \to 0,$$

by Chebyshev's inequality. It is easy to see that

$$\begin{aligned}
\mathbb{E}_{X,Z}\left[\frac{\beta(X_i)}{f_0(X_i)}K_\sigma(X_i - X)Z\right] &= \mathbb{E}_X\left[\frac{\frac{\partial G_{\theta,\rho}(X_i,h_0)}{\partial h}^\top}{f_0(X_i)}K_\sigma(X_i - X)\left[h_0(X) - h_0(X_i)\right]\right] \\
&= \frac{\frac{\partial G_{\theta,\rho}(X_i,h_0)}{\partial h}^\top}{f_0(X_i)}\left[\mathbb{E}_X(K_\sigma(X_i - X)h_0(X)) - h_0(X_i)\right],
\end{aligned}$$

and clearly $\mathbb{E}_X(K_\sigma(X_i - X)h_0(X)) \to h_0(X_i)$ as σ converges to zero. ■

Let $m_{n1}(z) = \int m_n(z, \tilde{z})\mathrm{d}F_0(\tilde{z}), m_{n2}(z) = \int m_n(\tilde{z}, z)\mathrm{d}F_0(\tilde{z})$, and $\mu = \int m_n(z, \tilde{z})\mathrm{d}F_0(\tilde{z})\mathrm{d}F_0(z)$. The following lemma is taken from Newey (1994).

Lemma 6 (V-Statistic Convergence) *If $Z_1, Z_2, \dots$ are i.i.d. then*

$$n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} m_n\left(Z_i, Z_j\right) - n^{-1}\sum_{i=1}^{n}\left[m_{n1}\left(Z_i\right) + m_{n2}\left(Z_i\right)\right] + \mu$$
$$= O_P\left\{\frac{\mathbb{E}\left[\left\|m_n\left(Z_1, Z_1\right)\right\|\right]}{n} + \frac{\left(\mathbb{E}\left[\left\|m_n\left(Z_1, Z_2\right)\right\|^2\right]\right)^{1/2}}{n}\right\}.$$

Proof of Proposition 11

We are interested in deriving the asymptotic distribution of

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} G_{\theta,\rho}(X_i, \hat{h}_0) - \mathbb{E}\left[G_{\theta,\rho}(X_i, h_0)\right]\right).$$

Our proof proceeds in two steps. First, we exploit results in Ackerberg et al. (2014), which allow us to recover the form of the adjustment term due to a noisy estimate of h_0. Second, we build upon Newey (1994) to derive the asymptotic distribution.

Step 1: (Derivation of the adjustment term $\alpha(W_i)$). $W_i = (X_i, Y_i)$. In comparison to the more general case treated in Ackerberg et al. (2014), the derivation of the adjustment term simplifies because $G_{\theta,\rho}(X_i, h_0)$ depends on h_0 only through the values it takes $h_0(X_i)$. For the sake of thoroughness, we follow and reproduce some developments and borrow some of their notations. For more details, we refer the interested reader to Ackerberg et al. (2014). Under H_0,

$$\mathbb{E}[G_{\theta,\rho}(X_i, h_0)] \geq 0,$$

and $h_0 = (h_{0,1}, ..., h_{0,L})$ consist of $L = \mathrm{card}(\mathcal{Y}) - 1$ nuisance functions, which are identified by the following conditional moment restrictions:

$$\mathbb{E}\left[\rho_l(W_i, h_{0,l}(X_i))|X_i\right] = 0 \; a.s \; X_i \quad \text{for } l = 1, ...L$$

where $\rho_l(W_i, h_{0,l}(X_i)) = 1\{Y_i = y_l\} - h_{0,l}(X_i)$. The last function is directly inferred from $\sum_{\mathcal{Y}} 1\{Y_i = y_l\} = 1$ almost surely. Every nuisance function h_l is assumed to lie in the space $\mathcal{H}_l$, which is a linear subspace of the space of integrable functions with respect to X. Finally, we define

$$k_l(X_i, h_{0,l}) = \mathbb{E}\left[\rho_l(W_i, h_{0,l}(X_i))|X_i\right].$$

One can easily show that the regularity conditions (p 922) of Ackerberg et al. (2014) are satisfied. In our case, k depends only on h through $h(X_i)$. Thus, the pathwise derivative with respect to h_l evaluated at h_0 in the direction $v_l \in \mathcal{V}_l \equiv \mathcal{H}_l - \{h_{0,l}\}$

$$\frac{\partial k_l(X_i, \theta, h_{0,l})}{\partial h_l}[v_l] = \frac{\partial \mathbb{E}[\rho_l(W_i, h_{0,l}(X_i) + \tau v_l(X_i))}{\partial \tau}\bigg|_{\tau=0} = \frac{\mathbb{E}\left[\rho_l(W_i, h_{0,l}(X_i))|X_i\right]}{\partial h_l} v_l(X) = -v_l(X).$$

with the second derivative being the ordinary derivative. Following Ackerberg et al. (2014) the Riesz representation theorem implies that there is a unique $u_l^* \in \mathcal{V}_l$ (set of measureable functions of X) such that $\forall v_l \in \mathcal{V}_l$,

$$\frac{\partial \mathbb{E}[G_{\theta,\rho}(X_i, h_0)]}{\partial h_l}[v_l] = \mathbb{E}\left[u_l^*(X_i) v_l(X_i)\right] \tag{A.9}$$

Now let us show that the adjustment term is equal to $\alpha(W_i) = \sum_{l=1}^{L} u_l^*(X_i)\rho_l(W_i, h_0(X_i))$. To do this, we follow the same reasoning as the one developed for the proof of proposition 1 in Newey (1994). Without loss of generality, we assume $L = 1$. As shown in Newey (1994), the adjustment term of the sum corresponds to the sum of the adjustment terms.

Consider a path $\{F_\tau(w)\}$ of the distribution of random variable W. Let h_τ be the function indexed by τ such that $\mathbb{E}_\tau\left[\rho(W_i, h_\tau(X_i))|X\right] = 0$ where $\mathbb{E}_\tau[.|X]$ denotes the conditional expectation taken under $F_\tau(w)$ with the corresponding score $S(w)$. From the definition of $u^*(X_i)$:

$$\frac{\partial}{\partial \tau}\mathbb{E}[G_{\theta,\rho}(X_i, h_\tau)] = \mathbb{E}\left[u^*(X_i)\frac{\partial}{\partial \tau}k(X_i, h_\tau)\right] = \frac{\partial}{\partial \tau}\mathbb{E}\left[u^*(X_i)k(X_i, h_\tau)\right].$$

Now by assumption on h_τ, we have that for any square integrable function $w(X_i)$,

$$\mathbb{E}_\tau\left[w(X_i)\rho(W_i, h_\tau(X_i))\right] = 0.$$

By differentiating with respect to τ, we have:

$$\frac{\partial}{\partial \tau}\mathbb{E}_\tau[w(X_i)\rho(W_i, h_0(X_i))] + \frac{\partial}{\partial \tau}\mathbb{E}\left[w(X_i)k(X_i, h_\tau)\right] = 0.$$

By combining the 2 previous equations, we have:

$$\frac{\partial}{\partial \tau}\mathbb{E}[G_{\theta,\rho}(X_i, h_\tau)] = -\frac{\partial}{\partial \tau}\mathbb{E}_\tau\left[u^*(X_i)\rho(W_i, h_0(X_i))\right] = \mathbb{E}_\tau\left[-u^*(X_i)\rho(W_i, h_0(X_i))S(W_i)\right].$$

Following equation (3.9) in Newey (1994), we have that the adjustment term writes $\alpha(W_i) = -u^*(X_i)\rho(W_i, h_0(X_i))$. Observe that we also have have that $\mathbb{E}[\alpha(W_i)] = 0$. Now let us derive $u^*(X_i)$ by exploiting the structure of our model. The same applies for the pathwise derivative of m, which depends only on h through $h(X_i)$:

$$\frac{\partial \mathbb{E}[G_{\theta,\rho}(X_i, h_0)|X_i]}{\partial h}[v] = \frac{\partial G_{\theta,\rho}(X_i, h_0)}{\partial h}[v] = \frac{\partial G_{\theta,\rho}(X_i, h_0)}{\partial h}v(X_i).$$

From the Riesz representation equation A.9, we have that for any v,

$$\mathbb{E}\left[\frac{\partial \mathbb{E}[G_{\theta,\rho}(X_i, h_0)|X_i]}{\partial h}[v] - u^*(X_i)v(X_i)\right] = \mathbb{E}\left[\left(\frac{\partial G_{\theta,\rho}(X_i, h_0)}{\partial h} - u^*(X_i)\right)v(X_i)\right]$$

The last equality holds for any v and in particular for $v(X_i) = \frac{\partial G_{\theta,\rho}(X_i, h_0)}{\partial h} - u^*(X_i)$. Thus, we have:

$$\mathbb{E}\left[\left(\frac{\partial G_{\theta,\rho}(X_i, h_0)}{\partial h} - u^*(X_i)\right)^2\right] = 0$$

Thus, $u^*(X_i) = \frac{\partial G_{\theta,\rho}(X_i, h_0)}{\partial h}$ almost surely. This yields the result:

$$\alpha(W_i) = \frac{\partial G_{\theta,\rho}(X_i, h_0)}{\partial h}^\top (\mathbf{1}\{Y_i = y\} - h_0(X_i)).$$

Step 2: (Derivation of the asymptotic distribution). By the mean-value Taylor expansion of $G_{\theta,\rho}(X_i, \hat{h}_0)$ around $G_{\theta,\rho}(X_i, h_0)$, there exists $\tilde{h} \in \mathcal{H}$ such that:

$$
\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n} G_{\theta,\rho}(X_i,\hat{h}_0)\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} G_{\theta,\rho}(X_i,h_0) + \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial G_{\theta,\rho}(X_i,h_0)}{\partial h}^{\top}(\hat{h}_0(X_i)-h_0(X_i))\\
&+ \underbrace{\frac{\rho}{\sqrt{n}}\sum_{i=1}^{n}(\hat{h}_0(X_i)-h_0(X_i))^{\top} J_\rho(X_i,\tilde{h})(\hat{h}_0(X_i)-h_0(X_i))}_{O_p(\rho n^{1/2-2\gamma})\ by\ 4(i)}\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} G_{\theta,\rho}(X_i,h_0) + \underbrace{\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial G_{\theta,\rho}(X_i,h_0)}{\partial h}^{\top}(\hat{h}_0(X_i)-h_0(X_i)) - \int \frac{\partial G_{\theta,\rho}(x,h_0)}{\partial h}^{\top}(\hat{h}_0(x)-h_0(x))\mathrm{d}x\right]}_{o_p(1)\text{ by Lemma } 5}\\
&+ \int \frac{\partial G_{\theta,\rho}(x,h_0)}{\partial h}^{\top}(\hat{h}_0(x)-h_0(x))\mathrm{d}x + \frac{\rho}{\sqrt{n}}\sum_{i=1}^{n}(\hat{h}_0(X_i)-h_0(X_i))^{\top} J_\rho(X_i,\tilde{h})(\hat{h}_0(X_i)-h_0(X_i))\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[G_{\theta,\rho}(X_i,h_0)+\alpha(W_i)\right] + \frac{\rho}{\sqrt{n}}\sum_{i=1}^{n}(\hat{h}_0(X_i)-h_0(X_i))^{\top} J_\rho(X_i,\tilde{h})(\hat{h}_0(X_i)-h_0(X_i))\\
&+ \underbrace{\sqrt{n}\left[\int \frac{\partial G_{\theta,\rho}(x,h_0)}{\partial h}^{\top}(\hat{h}_0(x)-h_0(x))\mathrm{d}x - \frac{1}{n}\sum_{i=1}^{n}\alpha(W_i)\right]}_{o_P(1)\text{ by Lemma } 4} + o_p(1)
\end{aligned}
$$

■

Proof of Proposition 12

Step 1: By the mean-value Taylor expansion of $G_{\theta,\rho_n}(X_i,\hat{h}_0)$ around $G_{\theta,\rho_n}(X_i,h_0)$, there exists $\tilde{h}\in\mathcal{H}$ such that:

$$
\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n} G_{\theta,\rho_n}(X_i,\hat{h}_0)\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} G_{\theta,\rho_n}(X_i,h_0) + \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial G_{\theta,\rho_n}(X_i,h_0)}{\partial h}^{\top}(\hat{h}_0(X_i)-h_0(X_i))\\
&+ \underbrace{\frac{\rho_n}{\sqrt{n}}\sum_{i=1}^{n}(\hat{h}_0(X_i)-h_0(X_i))^{\top}J_{\rho_n}(X_i,\tilde{h})(\hat{h}_0(X_i)-h_0(X_i))}_{O_p(\rho_n n^{1/2-2\gamma})\ by\ 4(i)}\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} G_{\theta,\rho_n}(X_i,h_0) + \sqrt{n}\underbrace{\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial G_{\theta,\rho_n}(X_i,h_0)}{\partial h}^{\top}(\hat{h}_0(X_i)-h_0(X_i)) - \int \frac{\partial G_{\theta,\rho_n}(x,h_0)}{\partial h}^{\top}(\hat{h}_0(x)-h_0(x))\mathrm{d}\right.}_{o_p(1) \text{ uniformly over } \rho_n \text{ by Lemma } 5}\\
&+ \int \frac{\partial G_{\theta,\rho_n}(x,h_0)}{\partial h}^{\top}(\hat{h}_0(x)-h_0(x))\mathrm{d}x + \frac{\rho_n}{\sqrt{n}}\sum_{i=1}^{n}(\hat{h}_0(X_i)-h_0(X_i))^{\top}J_{\rho_n}(X_i,\tilde{h})(\hat{h}_0(X_i)-h_0(X_i))\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[G_{\theta,\rho_n}(X_i,h_0)+\alpha(W_i)\right] + \frac{\rho_n}{\sqrt{n}}\sum_{i=1}^{n}(\hat{h}_0(X_i)-h_0(X_i))^{\top}J_{\rho_n}(X_i,\tilde{h})(\hat{h}_0(X_i)-h_0(X_i))\\
&+ \underbrace{\sqrt{n}\left[\int \frac{\partial G_{\theta,\rho_n}(x,h_0)}{\partial h}^{\top}(\hat{h}_0(x)-h_0(x))\mathrm{d}x - \frac{1}{n}\sum_{i=1}^{n}\alpha(W_i)\right]}_{o_P(1) \text{ uniformly over } \rho_n \text{ by Lemma } 4} + o_p(1)
\end{aligned}
$$

with $\alpha(W_i) = \frac{\partial G_{\theta,\rho_n}(X_i,h_0)}{\partial h}^{\top}(\mathbf{1}\{Y_i=y\}-h_0(X_i))$, $J_{\rho_n}(x_i,\tilde{h})$ such that $\frac{\partial^2 G_{\theta,\rho_n}(x_i,\tilde{h})}{\partial h \partial h^{\top}} = \rho J_{\rho}(x_i,\tilde{h})$ and for any $(x_i,\tilde{h})$, $||J_{\rho_n}(x_i,\tilde{h})||_{\infty} \leq 2m$. So, we have

$$
\begin{aligned}
\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} G_{\theta,\rho_n}(X_i,\hat{h}_0) - \mathbb{E}\left[G_{\theta,\rho_n}(X_i,h_0)\right]\right) &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} G_{\theta,\rho_n}(X_i,h_0) + \alpha_n(W_i) - \mathbb{E}\left[G_{\theta,\rho_n}(X_i,h_0)\right]\right)\\
&+ O_p(\rho n^{1/2-2\gamma}) + o_p(1).
\end{aligned}
$$

Step 2: From triangular array CLT, we have

$$
\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} G_{\theta,\rho_n}(X_i,h_0) + \alpha_n(W_i) - \mathbb{E}\left[G_{\theta,\rho_n}(X_i,h_0)\right]\right) \xrightarrow{d} \mathcal{N}(0,V_0),
$$

with $V_0 = \lim_{n\to\infty} \mathrm{Var}\left[\left[G_{\theta,\rho_n}(X_i,h_0)+\alpha_n(W_i)\right]\right]$ as long as Lyapunov condition is satisfies or

$$
\frac{1}{n^{\frac{\delta}{2}}}\mathbb{E}\left[\frac{|G_{\theta,\rho_n}(X_i,h_0)+\alpha_n(W_i)-\mathbb{E}\left[G_{\theta,\rho_n}(X_i,h_0)\right]|^{2+\delta}}{\mathrm{Var}(G_{\theta,\rho_n}(X_i,h_0)+\alpha_n(W_i))^{1+\frac{\delta}{2}}}\right] \to 0,
$$

for some $\delta > 0$. It is easy to check that this holds as

$$\begin{aligned}G_{\theta,\rho_n}(X_i,h_0)+\alpha_n(W_i) &= -\rho^{-1}\log\Big(\frac{1+\sum_{j=1}^{p}\exp(\rho(q_j^\top h(X_i)-C_{\theta,j}))}{1+p}\Big) \\ &\quad + \frac{\partial G_{\theta,\rho_n}(X_i,h_0)}{\partial h}^\top(\mathbf{1}\{Y_i=y\}-h_0(X_i)) \\ &\leq \frac{\log(p+1)}{\rho}+2.\end{aligned}$$

Similarly, we can show that it is also bounded below. Since $|G_{\theta,\rho_n}(X_i,h_0)+\alpha_n(W_i)|$ is bounded, Lyapunov condition is automatically satisfied.

Step 3: Let V_n be a consistent estimator for V_0. Now, we show asymptotic validity and consistency. Under H_0, $\theta \in \Theta_I$,

$$\theta \in \Theta_I \implies G_{\theta,\rho_n}(X_i,h_0) \geq 0 \ \text{ a.s } \implies \mathbb{E}\left[G_{\theta,\rho_n}(X_i,h_0)\right] \geq 0 \ \forall n.$$

(i) Asymptotic validity:

$$\begin{aligned}&\Pr\left(\sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^{n}G_{\theta,\rho_n}(X_i,\hat{h}_0)}{\sqrt{V_n}} \geq z_\alpha\right) \\ =&\Pr\left(\sqrt{n}\frac{\mathbb{E}\left[G_{\theta,\rho_n}(X_i,h_0)\right]}{\sqrt{V_n}} + \sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^{n}G_{\theta,\rho_n}(X_i,\hat{h}_0)-\mathbb{E}\left[G_{\theta,\rho_n}(X_i,h_0)\right]}{\sqrt{V_n}} \geq z_\alpha\right) \\ \geq&\Pr\left(\sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^{n}G_{\theta,\rho_n}(X_i,\hat{h}_0)-\mathbb{E}\left[G_{\theta,\rho_n}(X_i,h_0)\right]}{\sqrt{V_n}} \geq z_\alpha\right) \\ =&\Pr\left(\sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^{n}G_{\theta,\rho}(X_i,h_0)+\alpha(W_i)-\mathbb{E}\left[G_{\theta,\rho}(X_i,h_0)\right]+o_P(1)}{\sqrt{V_n}} \geq z_\alpha\right) \qquad \underset{n\to\infty}{\rightarrow} 1-\alpha\end{aligned}$$

(ii) Consistency:

$$\begin{aligned}\theta \notin \Theta_I &\implies \mathbb{E}\left[\lim_{n\to\infty}G_{\theta,\rho_n}(X_i,h_0)\right] = \gamma < 0 \\ &\implies \frac{1}{n}\sum_{i=1}^{n}G_{\theta,\rho_n}(X_i,\hat{h}_0) \xrightarrow{P} \gamma \\ &\implies \frac{\frac{1}{n}\sum_{i=1}^{n}G_{\theta,\rho_n}(X_i,\hat{h}_0)}{\sqrt{V_n}} - \frac{\gamma}{\sqrt{V_0}} \xrightarrow{P} 0\end{aligned}$$

Second line: from the asymptotic distribution in step 2: $\frac{1}{n}\sum_{i=1}^{n}G_{\theta,\rho_n}(X_i,\hat{h}_0) = \mathbb{E}\left[G_{\theta,\rho_n}(X_i,h_0)\right] +$

$o_p(1)$. Thus,

$$\begin{aligned}
&\Pr\left(\sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^{n}G_{\theta,\rho_n}(X_i,\hat{h}_0)}{\sqrt{V_n}}\geq z_\alpha\right)\\
&=\Pr\left(\frac{\frac{1}{n}\sum_{i=1}^{n}G_{\theta,\rho_n}(X_i,\hat{h}_0)}{\sqrt{V_n}}-\frac{z_\alpha}{\sqrt{n}}-\frac{\gamma}{\sqrt{V_0}}\geq-\frac{\gamma}{\sqrt{V_0}}\right)\\
&\leq\Pr\left(\left|\frac{\frac{1}{n}\sum_{i=1}^{n}G_{\theta,\rho_n}(X_i,\hat{h}_0)+\log(p)/\rho_n}{\sqrt{V_n}}-\frac{z_\alpha}{\sqrt{n}}-\frac{\gamma}{\sqrt{V_0}}\right|\geq-\frac{\gamma}{\sqrt{V_0}}\right)\underset{n\to\infty}{\to}0.
\end{aligned}$$

■

Proof of Proposition 13

Step 1: Denote

$$R_{n,1}:=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[\tilde{G}_{\theta,\rho}(X,\hat{h}_{-k})\right]-\frac{1}{n}\sum_{i=1}^{n}\tilde{G}_{\theta,\rho}(X_i,h_0),$$

where $\mathbb{E}_{n,k}(\phi(X))=\frac{K}{n}\sum_{i\in S_k}\phi(X_i)$. In step (2), we will show that

$$R_{n,1}=O_P\left(r_n\right).$$

We have

$$\hat{\alpha}=\frac{1}{K}\sum_{k=1}^{K}\frac{K}{n}\sum_{i\in S_k}\tilde{G}_{\theta,\rho}(X_i,\hat{h}_{-k}(X_i))=\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[\tilde{G}_{\theta,\rho}(X,\hat{h}_{-k})\right],$$

and

$$\begin{aligned}
\sqrt{n}\left(\hat{\alpha}-\alpha\right)&=\sqrt{n}\left(\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{n,k}\left[\tilde{G}_{\theta,\rho}(X,\hat{h}_{-k})\right]-\alpha\right)\\
&=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\tilde{G}_{\theta,\rho}(X_i,h_0)-\mathbb{E}(\tilde{G}_{\theta,\rho}(X_i,h_0))\right]+\sqrt{n}R_{n,1}.
\end{aligned}$$

Moreover, because $\sqrt{n}r_n=o(1)$, it follows that

$$\sqrt{n}\left(\hat{\alpha}-\alpha\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\tilde{G}_{\theta,\rho}(X_i,h_0)-\mathbb{E}(\tilde{G}_{\theta,\rho}(X_i,h_0))\right]+o\left(1\right),$$

Combining with the Lindeberg-Feller central limit theorem yields the result.

Step 2: In this step, we establish bound on $R_{n,1}$. Again, because K is a fixed integer, which is independent of n, it suffices to show that for any k,

$$\mathbb{E}_{n,k}\left[\tilde{G}_{\theta,\rho}(X_i,\hat{h}_{-k})\right]-\mathbb{E}_{n,k}\left[\tilde{G}_{\theta,\rho}(X_i,h_0)\right]=O_P\left(r_n\right).$$

To do so, fix any k and introduce the following additional empirical process notation,

$$\mathbb{G}_{n,k}[\phi(W)] = \sqrt{\frac{k}{n}} \sum_{i \in S_k} \left(\phi\left(W_i\right) - \int \phi(w) dP \right),$$

where ϕ is any P-integrable function on $\mathcal{W}$. Then observe that by the triangle inequality,

$$\left\| \mathbb{E}_{n,k}\left[\tilde{G}_{\theta,\rho}(X_i, \hat{h}_{-k})\right] - \mathbb{E}_{n,k}\left[\tilde{G}_{\theta,\rho}(X_i, h_0)\right] \right\| \leq \sqrt{\frac{k}{n}} \left[\mathcal{I}_{1,k} + \mathcal{I}_{2,k}\right],$$

where for $W_i = (X_i, Y_i)$,

$$\begin{aligned}
\mathcal{I}_{1,k} &:= \left\| \mathbb{G}_{n,k}\left[\tilde{G}_{\theta,\rho}(X_i, \hat{h}_{-k})\right] - \mathbb{G}_{n,k}\left[\tilde{G}_{\theta,\rho}(X_i, h_0)\right] \right\|, \\
\mathcal{I}_{2,k} &:= \sqrt{\frac{n}{k}} \left\| \mathbb{E}_P\left[\tilde{G}_{\theta,\rho}(X_i, \hat{h}_{-k}) \mid (W_i)_{i \in S_{-k}}\right] - \mathbb{E}_P\left[\tilde{G}_{\theta,\rho}(X_i, h_0)\right] \right\|.
\end{aligned}$$

To bound $\mathcal{I}_{1,k}$, note that, as above, conditional on $(W_i)_{i \in S_{-k}}$, the estimator $\hat{h}_{-k}$ is non-stochastic, and so we have,

$$\begin{aligned}
\mathbb{E}_P\left[\mathcal{I}_{1,k}^2 \mid (W_i)_{i \in l_k^c}\right] &= \mathbb{E}_P\left[\left\|\tilde{G}_{\theta,\rho}(X_i, \hat{h}_{-k}) - \tilde{G}_{\theta,\rho}(X_i, h_0)\right\|^2 \,\middle|\, (W_i)_{i \in S_{-k}}\right] \\
&\leq \sup_{\|h-h_0\| = O_p(n^{-\gamma})} \mathbb{E}_P\left[\left\|\tilde{G}_{\theta,\rho}(X_i, h) - \tilde{G}_{\theta,\rho}(X_i, h_0)\right\|^2 \,\middle|\, (W_i)_{i \in l_k^c}\right] \\
&\leq \sup_{\|h-h_0\| = O_p(n^{-\gamma})} \mathbb{E}_P\left\|\tilde{G}_{\theta,\rho}(X_i, h) - \tilde{G}_{\theta,\rho}(X_i, h_0)\right\|^2 \\
&= \sup_{\|h-h_0\| = O_p(n^{-\gamma})} \mathbb{E}_P\left\|G(X_i, h) - G(X_i, h_0)\right\|^2 \\
&\quad + \sup_{\|h-h_0\| = O_p(n^{-\gamma})} \mathbb{E}\left\|\frac{\partial G(h)}{\partial h}^\top (\mathbf{1}\{Y_i = \mathbf{y}\} - h(X_i)) - \frac{\partial G(h_0)}{\partial h}^\top (\mathbf{1}\{Y_i = \mathbf{y}\} - h_0(X_i))\right\|^2 \\
&\leq \sup_{\|h-h_0\| = O_p(n^{-\gamma})} \mathbb{E}\left\|\frac{\partial G_\rho(\overline{h})}{\partial h}^\top (h(X_i) - h_0(X_i))\right\|^2 \\
&\quad + \sup_{\|h-h_0\| = O_p(n^{-\gamma})} \mathbb{E}\left\|\frac{\partial G_\rho(h_0)}{\partial h}^\top (h(X_i) - h_0(X_i))\right\|^2 \\
&\quad + \sup_{\|h-h_0\| = O_p(n^{-\gamma})} \mathbb{E}\left\|\left(\frac{\partial^2 G_\rho(\tilde{h})}{\partial h \partial h^\top} (\mathbf{1}\{Y_i = \mathbf{y}\} - h(X_i))\right) (h(X_i) - h_0(X_i))\right\|^2 \\
&= o(n^{-2\gamma}).
\end{aligned}$$

To bound $\mathcal{I}_{2,k}$, introduce the function

$$f_k(r) := \mathbb{E}_P\left[\tilde{G}\left(X_i, h_0 + r\left(\hat{h}_{-k} - h_0\right)\right) \,\middle|\, (W_i)_{i \in S_k}\right] - \mathbb{E}_P\left[\tilde{G}_{\theta,\rho}(X_i, h_0)\right], \quad r \in [0, 1].$$

Then, by Taylor expansion,

$$f_k(1) = f_k(0) + f_k'(0) + f_k''(\tilde{r})/2, \quad \text{for some } \tilde{r} \in (0,1).$$

However, $\|f_k(0)\| = 0$ because

$$\mathbb{E}_P\left[\tilde{G}_{\theta,\rho}(X_i, h_0) \mid (W_i)_{i \in S_{-k}}\right] = \mathbb{E}_P\left[\tilde{G}_{\theta,\rho}(X_i, h_0)\right].$$

In addition, from the orthogonality condition

$$f_k'(0) = 0.$$

Moreover, we have,

$$\begin{aligned}
\|f_k''(\tilde{r})\| &\leq \sup_{r\in(0,1)} \|f_k''(r)\| \\
&\leq \sup_{r\in(0,1)} \sup_{\|h-h_0\|=O_p(n^{-\gamma})} \mathbb{E}\left[(h(X_i)-h_0(X_i))^\top \frac{\partial^2 G(X_i,\tilde{r})}{\partial h \partial h^\top}(h(X_i)-h_0(X_i))\right] \\
&= o(n^{-2\gamma}).
\end{aligned}$$

■