Testing Information Ordering for Strategic Agents*

Sukjin Han
School of Economics
University of Bristol
vincent.han@bristol.ac.uk

Hiroaki Kaido
Department of Economics
Boston University
hkaido@bu.edu

Lorenzo Magnolfi
Department of Economics
University of Wisconsin
magnolfi@wisc.edu

March 2025

Abstract

A key primitive of a strategic environment is the information available to players. Specifying a priori an information structure is often difficult for empirical researchers. We develop a test of *information ordering* that allows researchers to examine if the true information structure is at least as informative as a proposed baseline. We construct a computationally tractable test statistic based on moment inequalities by utilizing the notion of Bayes Correlated Equilibrium (BCE) to translate the ordering of information structures into an ordering of functions. We apply our test to examine whether hubs provide informational advantages to certain airlines in addition to market power.

Keywords: Information Structure, Bayes Correlated Equilibria, Semiparametric Tests.

*We thank Camilla Roncoroni for inputs on an earlier version of this manuscript. We appreciate Gaurab Aryal, Debopam Bhattacharya, Jörgen Weibull and participants at the Cemmap Workshop on Econometrics, Models of Strategic Interactions, and SWERIE 2025 for useful feedback. Kaido gratefully acknowledges financial support from NSF grants SES-1824344 and SES-2018498.

1 Introduction

Many important economic interactions are strategic in nature. When analyzing data generated in such contexts, researchers often bring to data models of games to estimate primitives and perform counterfactual simulations. Examples include the analysis of firms' entry decisions and bidding behavior in auctions. One such key primitive is the information structure, i.e. a full description of the information available to players as they interact and generate the data. Information plays an important role in shaping outcomes: for example, potential entrants may receive a common signal about the unobserved profitability of a market, or certain bidders may know more than others about the realized valuations of bidders in an auction. Therefore, the information structure may be either of independent economic interest—e.g., because asymmetries in information may suggest that some firms have political connections (Magnolfi and Roncoroni, 2016a; Baltrunaite, 2020) or corruption—or crucial in evaluating counterfactual policy—as misspecifying the information structure leads to misleading predictions.

The information structure that prevails in the strategic interaction is seldom known to the researcher. Thus, the researcher may need to infer information based on data on observable outcomes. However, the estimation of information structure is typically challenging. First, the space of possible information structures is potentially very large, as it encompasses any form of signals, with different degrees of informativeness, that players may receive. Restrictive parametrizations of information structures are thus likely to result in misspecification. Treating the information structure as a nonparametric object would be an alternative, but the high dimensionality of this object makes it unfeasible with standard data used in applied research.

To make progress on this problem, we focus on conducting inference on the information structure. We consider models of discrete games, and adopt the orderings of information structures defined in Bergemann and Morris (2016). Their results allow us to compare information structures and establish a precise notion of the *ordering* of information structures. Crucially, their results also imply that different levels of information have distinct implications for the observables, thus providing a basis for testing. We can thus formulate statistical hypotheses to test whether the information structure that prevails in the data exceeds a certain baseline.

Suppose the players have access to the information described by structure $S(x)$ in a game characterized by observable characteristics x. Roughly, the structure describes the precision of the signal each player receives about the underlying payoff state. The players then play a Bayes-Nash equilibrium (BNE) strategy profile under $S(x)$. The researcher observes the equilibrium play. While $S(x)$ is unobservable, the researcher postulates that it is as informative as a certain baseline information structure $S^r(x)$. The baseline structure describes the minimal information the researcher believes the players can access. Specifying $S^r(x)$ is typically easier than specifying $S(x)$ because the former describes a minimal information structure, while the latter requires the precise structure of the players' true knowledge. Bergemann and Morris (2016) show that, for a specific notion of ordering called *individual sufficiency*, one can translate the ordering of information into that of the equilibrium predictions. Their solution concept *Bayes Correlated Equilibrium (BCE)* is useful for two main reasons. First, it is a general solution concept that encompasses BNE in terms of prediction. If data are generated from a BNE under $S(x)$, its *conditional choice probability* belongs to the set of BCE conditional choice probabilities with baseline information structure $S^r(x)$. Second, characterizing the set of BCE conditional choice probabilities, the *BCE prediction*, is computationally tractable. This is because the BCE prediction is a convex set characterized by linear inequalities and equalities.

We use the properties above to construct a test statistic for the null hypothesis that $S(x)$ that generates the conditional choice probability is as informative as $S^r(x)$. Our test statistic is based on the distance from the empirical conditional choice probability to the BCE prediction under the baseline information structure. This construction exploits the isometry between convex sets and their support functions. The null hypothesis is rejected when the test statistic exceeds a bootstrap critical value. The critical value is calculated so that the test is asymptotically valid uniformly over a large class of data-generating processes. According to our preliminary Monte Carlo experiments, the proposed test properly controls its size and can detect violations of the restrictions imposed by the null hypothesis. Specifically, in an incomplete information game, we tested the null hypothesis that some payoff shocks are known to all players. The test showed monotonically increasing power when the conditional choice probabilities deviate locally from the BCE prediction under the null hypothesis. We also extend our approach to testing a sequence of hypotheses, which can help refine the understanding of a game's information structure.

Our testing approach complements recent approaches to estimation and counterfactual simulation that maintain weak assumptions on information, either using a moment inequalities approach (Dickstein and Morales, 2018), or using BCE (Magnolfi and Roncoroni, 2022; Syrgkanis et al., 2017; Gualdani and Sinha, 2023). Similar to this latter literature, we rely on BCE as a solution concept and build on the theoretical literature that established its desirable properties (Bergemann and Morris, 2016). Whereas the estimation literature focuses on recovering the game's payoffs while being agnostic about information, we pursue testing of hypotheses on the information structure, which is useful whenever the information structure is an object of interest.

One of the main practical hurdles when performing estimation under weak assumptions on information is that counterfactual prediction may be uninformative. Our testing approach, similar to other recent methods (Bergemann et al., 2017), may help produce sharper counterfactuals. When our test rejects null hypotheses in favor of a more informative information structure, which can be used in counterfactuals, the researcher obtains a tighter set of predictions.

In work-in-progress, we apply our test to examine whether hubs provide informational advantages to certain airlines in addition to market power. Beyond our application, there are a range of strategic empirical contexts where inference on the information structure is of economic interest. For instance, Magnolfi and Roncoroni (2016a) study entry in the Italian supermarket industry, where one player has political connection that can both affect payoffs directly, and affect information—e.g., by providing the connected player with information about rivals' costs. Similarly, Baltrunaite (2020) studies procurement auctions in which firms can buy preferential treatment via donations to politicians. Preferential treatment in this context seems to arise via an informational channel, where bidders receive information on competitors' valuations or bids. Our test can be applied to similar environments to, e.g., detect the presence of political connections or corruption.

There is a rich literature on the econometric analysis of games (see de Paula, 2013, for a summary). A common empirical practice is to specify a particular information structure and estimate the model parameters. Earlier examples of testing for the information structure in the literature include Navarro and Takahashi (2012) and Grieco (2014). The latter is based on a parametrization of the information structure that nests incomplete and complete information but is strictly less general than our model.The games above make incomplete (or set-valued) predictions. That is, given observed and unobserved exogenous variables, the model admits multiple solutions (Jovanovic, 1989). There have been developments on ways to systematically derive identifying restrictions in such models without making assumptions on unknown (equilibrium) selection mechanisms (Tamer, 2003; Beresteanu et al., 2011; Galichon and Henry, 2011). Our approach also leverages robust implications of a specific incomplete model that admits equilibria with unknown information structures. The implied restrictions exploit the convexity of the BCE prediction and take the form of moment inequalities. As such, we utilize the support function of the BCE prediction to construct a test statistic and use a bootstrap-based critical value. Statistics based on support functions are used in other models, including regression models with interval-valued data. We propose a new way to "studentize" the sample moments used to calculate the statistic. A well-known issue with inference based on moment inequalities is that test statistics may have different asymptotic distributions depending on the configuration of the data-generating process (see Canay and Shaikh, 2017, for a summary and reference there). We adopt a moment selection procedure as to ensure that the test is asymptotically valid uniformly across a wide range of data-generating processes.

2 Model

We describe a general model of an empirical discrete game, the environment where we develop our testing procedure. This model is similar to the one described in Magnolfi and Roncoroni (2022). Games in the class we describe are indexed by realizations of covariates $x \in X$. Players are indexed in a finite set $N \equiv \{1, ..., |N|\}$. Each player $i \in N$ chooses an action $y_i \in Y_i$, a discrete set. Both the actions' set $Y \equiv \times_{i \in N} Y_i$ and N do not depend on x. All these aspects of the game are common knowledge among players and known to the researcher. The researcher jointly observes actions $y \in Y$ and covariates x. Next, we describe payoff structure and information structure of the game.

2.1 Payoff Structure and Information Structure

Player i is characterized by a payoff type $\varepsilon_i \in \mathcal{E}_i$. The vector of payoff types $\varepsilon = (\varepsilon_i)_{i \in N}$ is distributed according to the cumulative distribution function (CDF) $F(\cdot; \theta_\varepsilon)$, which is known to the researcher up to the finite-dimensional vector θ_ε. Payoffs to player i, are denoted by π_i, and are realized according to the function $\pi_i(\cdot, \cdot; x, \theta_\pi) : Y \times \mathcal{E}_i \to \mathbb{R}$. We assume that payoff types ε are independent of covariates x. A realization of x and a vector of parameters $\theta = (\theta_\pi, \theta_\varepsilon) \in \Theta$ fully characterize the payoff structure of the game.

We assume that every player i knows the parameters and x. The players also observe a private random signal $\tau_i^x \in \mathcal{T}_i$, which may carry information on the vector of payoff types ε. The set $\mathcal{T}$ is the space of signals equipped with a σ-algebra $\mathcal{F}$.[1]

We define a generic information structure S as a mapping from values of covariates to conditional distributions of signals:

$$S : x \mapsto \left(\mathcal{T}_x, \left\{P^x_{\tau|\varepsilon} : \varepsilon \in \mathcal{E}\right\}\right),$$

where the probability kernel $\{P^x_{\tau|\varepsilon} : \varepsilon \in \mathcal{E}\}$ is the collection of probability distributions of τ^x, the vector of signals, conditional on every realization of ε. The set $\mathcal{T}_x \subset \mathcal{T}$ is the support of such distributions. Let

$$\mathbb{M} \equiv \{(\mathcal{T}', \{P_{\tau|\varepsilon} : \varepsilon \in \mathcal{E}\}) : \mathcal{T}' \subset \mathcal{T},\ P_{\tau|\varepsilon}(\mathcal{T}') = 1\ a.s.\}$$

[1]In what follows, we tacitly assume the signals are defined on the measurable space $(\mathcal{T}, \mathcal{F})$ but omit the underlying σ-algebra from the definition of information structure for notational simplicity.

be the collection of the conditional distributions of signals, and view the information structure $S : X \to \mathbb{M}$ as an unknown $\mathbb{M}$-valued nonparametric function. We allow information structure to depend on x because the informational environment may change with the observable covariates. In what follows, we use $\Gamma^x(\theta, S)$ to indicate the game with covariates x, payoff parameters θ and information structure $S(x)$.

Example 1: Consider the entry game first proposed in Bresnahan and Reiss (1991), where players are two firms that are potential entrants in a market, and choose to either "Enter" or "Not enter", corresponding to $y_i = 1$ and $y_i = 0$, respectively. The researcher observes entry choices across a set of markets with covariates x. Firm i's profits are zero when not entering, and $\pi_i(y, \varepsilon_i; x, \theta_\pi) = x'\beta + \Delta y_{-i} + \varepsilon_i$ upon entry. The payoff types (e.g., cost shifters) ε_i are unobservable to the researcher; an information structure S specifies the information that player i has on its opponent's ε_{-i}. In addition to knowing its own payoff type, firm i may have access to a noisy measurement of the opponent firms' unobserved cost shifters $\varepsilon_{-i} \in \mathbb{R}^{|N|-1}$. One may model the signal as a random vector $\tau_i^x \in \mathcal{T}^x = \mathbb{R}^{|N|}$ following an unknown conditional distribution $P^x_{\tau|\varepsilon}$.

We introduce here some special cases of information structure that are useful in what follows. We call the *null information* structure S_{Null} the one characterized by fully uninformative signals, so that each τ_i does not affect players' beliefs on the realization of ε. In terms of the conditional law of the signal, $P^x_{\tau|\varepsilon}$ corresponds to the setting where τ is independent of ε for all x. We define the *incomplete information* structure S_I as the one characterized by signals τ_i that reveal the realization of ε_i, but are only partially informative on ε_{-i}. Coupled with the assumption of independent types, this assumption is adopted in seminal work on social interaction and econometrics of games (e.g., Brock and Durlauf, 2001; Seim, 2006). Conversely, in a game with the *complete information* structure S_C, the signal τ_i is fully informative on the vector ε (as in e.g., Bresnahan and Reiss, 1991; Tamer, 2003). Finally, consider the *privileged information* structure S_P. To simplify the discussion, we define this information structure in the context of the two players game of Example 1; extensions to more general games are immediate. In S_P player 1 has more information than player 2. In this case, τ_2 is only informative on ε_2, whereas τ_1 fully reveals $\varepsilon = (\varepsilon_1, \varepsilon_2)$. Notice that, for notational simplicity, in describing the information structures above we omit the dependence on x of signal kernels and support sets. An information structure $S(x)$ could feature, for different values of x, any of the special information structures described above.

2.2 Comparing Information Structures

To make progress towards our goal of using the data to test hypotheses on the information structure of the game, we need to be able to have a rigorous way of comparing information structures. This is a complex task. Fix a value of x and consider, for instance, two information structures for the same collection of games:

$$\begin{aligned} S^1(x) &= \left(\mathcal{T}^{1,x}, \left\{P^{1,x}_{\tau|\varepsilon} : \varepsilon \in \mathcal{E}\right\}\right) \\ S^2(x) &= \left(\mathcal{T}^2_x, \left\{P^{2,x}_{\tau|\varepsilon} : \varepsilon \in \mathcal{E}\right\}\right) \end{aligned}$$

Each information structure can be a complex, high-dimensional object, so that it is not immediate to compare $S^1(x)$ and $S^2(x)$. More precisely, we would want to form an *ordering* among information structures, as to give a rigorous meaning to the statement "$S^1(x)$ is more informative than $S^2(x)$." We develop two interrelated notions of ordering in this subsection.

The first notion of ordering starts from a simple consideration: if there is an information structure that contains all the information present in $S^1(x)$ and $S^2(x)$, then this new combined information structure is clearly at least as informative as either $S^1(x)$ or $S^2(x)$. We thus follow Bergemann and Morris (2016) and define combinations of information structures. For this, we assume the space $\mathcal{T}$ is finite.

Definition 1 (Combination)**.** *The information structure at x*

$$S^*(x) = \left(\mathcal{T}^{*,x}, \left\{P^x_{\tau^*|\varepsilon} : \varepsilon \in \mathcal{E}\right\}\right)$$

is a combination of $S^1(x)$ and $S^2(x)$ if

$$\begin{aligned} \mathcal{T}^{*,x} &= \prod_{i=1}^{|N|} \mathcal{T}_i^{*,x} \\ \mathcal{T}_i^{*,x} &= \mathcal{T}_i^{1,x} \times \mathcal{T}_i^{2,x}, \quad i \in N. \\ \sum_{\tau^2} P^x_{\tau^*|\varepsilon}(\tau^1, \tau^2|\varepsilon) &= P^x_{\tau^1|\varepsilon}(\tau^1|\varepsilon) \quad \textit{for each } \tau^1, \\ \sum_{\tau^1} P^x_{\tau^*|\varepsilon}(\tau^1, \tau^2|\varepsilon) &= P^x_{\tau^2|\varepsilon}(\tau^2|\varepsilon) \quad \textit{for each } \tau^2. \end{aligned}$$

Intuitively, the combined information structure $S^*(x)$ gives players access to both $S^1(x)$ and $S^2(x)$. Based on the definition of combination, we can also define, for any information structure $S^1(x)$, its *expansions*.

Definition 2 (Expansion)**.** *An information structure $S^*(x)$ is an expansion of $S^1(x)$ if it is a combination of $S^1(x)$ and $S^2(x)$ for some $S^2(x)$. We write $S^* \succeq_{Exp} S^1$ if $S^*(x)$ is an expansion for $S^1(x)$ for all $x \in X$.*

Clearly, if $S^*(x)$ is an expansion of $S^1(x)$, it is more informative than $S^1(x)$ for the players that observe its signals. We consider settings in which players of a game are believed to have access to a baseline information structure. For this purpose, we define a set of information structures that are at least as informative as some baseline.

Definition 3 (Set of Expansions)**.** *For any $S_B : X \to \mathbb{M}$, $\mathcal{S}(S_B)$ is the set of all information structures that are expansions of S_B.*

When constructing $\mathcal{S}(S_B)$, we refer to S_B as the *baseline* information structure for the set—in fact, any information structure in $\mathcal{S}(S_B)$ will be at least as informative as S_B in the sense of expansions. Moreover, notice that in our context, every information structure is an expansion of the null information S_N. Thus, we define $\mathcal{S} = \mathcal{S}(S_N)$ as the universe of information structures that we consider. One may view $\mathcal{S}$ as the parameter space for the information structure.

We also consider a second notion of ordering of information structures to make further progress and use BCE predictions to compare information structures. Bergemann and Morris (2016) present the following intuitive way to (partially) order information structures:

Definition 4 (Individual Sufficiency)**.** *$S^1(x)$ is individually sufficient for $S^2(x)$ if there exist a combined information structure $S^*(x)$ such that, for each i and measurable set $A \subset \mathcal{T}_i$,*

$$P^{*,x}(\tau_i^2 \in A|\tau_i^1, \tau_{-i}^1, \varepsilon) = P^{*,x}(\tau_i^2 \in A|\tau_i^1), \ \textit{a.s.}$$

where the probability of $P^{*,x}(\tau_i^2|\tau_i^1, \tau_{-i}^1, \varepsilon)$ is computed using the combined kernel $P^{*,x}_{\tau^*|\varepsilon}$.

In short, $S^1(x)$ is individually sufficient for $S^2(x)$ if there is a combined information structure of the two, in which, for each i,

$$\tau_i^2 \perp (\tau_{-i}^1, \varepsilon)|\tau_i^1. \tag{1}$$

Hence, a player, given his signal τ_i^1, and given he knows the information structures $S^1(x)$ and $S^2(x)$, is able to compute what signal he would have received according to the information structure $S^2(x)$. The individual sufficiency property reduces to Blackwell sufficiency in the one-player case. Clearly, the ordering of information structures based on individual sufficiency is a partial one. Let us introduce the ordering of information structures based on the indivdual sufficiency.

Definition 5. *An information structure S^1 is individually sufficient for S^2 if $S^1(x)$ is individually sufficient for $S^2(x)$ for all $x \in X$. We write $S^1 \succeq S^2$ whenever S^1 is individually sufficient for S^2.*

Example 2: Consider again the two-player of entry game introduced in Example 1. Suppose that payoff types for each player i are $\varepsilon_i = \nu_i + \epsilon_i$. We define the *public signals* information structure S_{Pub} as one where signals τ_i reveal the opponent's shock ν_{-i} symmetrically for each player. Clearly, this information structure represents an expansion of S_I, and is individually sufficient for S_I so that we can write $S_{Pub} \succeq S_I$.

There is a tight relation between the partial orders of information structures induced by expansion and individual sufficiency. In fact, these two concepts are interchangeable if we consider as the same all the information structures that have the same canonical representation (Mertens and Zamir, 1985), or induce the same beliefs about the state. This assumption, which refines the space of all information structures, is natural in our context. Ultimately, we are interested in testing hypotheses on information using data, and it is not possible to distinguish empirically two information structures that induce the same beliefs, and hence the same actions. Hence, if we restrict our attention to information structures with different canonical representation, Claim 1 in Bergemann and Morris (2016) implies:

Lemma 1. *For any baseline information structure S_B, we let the set of information structures $\mathcal{S}^*(S_B)$ be the subset of $\mathcal{S}(S_B)$ such that any two $S, S' \in \mathcal{S}^*(S_B)$ have a different canonical representation. Then, for any $S, S' \in \mathcal{S}^*(S_B)$, $S \succeq S'$ if and only if $S \succeq_{Exp} S'$.*

The Lemma defines the set $\mathcal{S}^*(S_B)$ as the collection of all possible information structures S that have a different canonical representation and are expansions of the baseline S_B. Within such sets $\mathcal{S}^*(S_B)$, the partial orders implied by expansion and individual sufficiency are equivalent. We take $\mathcal{S}^* \equiv \mathcal{S}^*(S_N)$ as the parameter space for the unknown information structure.

2.3 Equilibrium Concepts and Predictions

We define two notions of equilibrium in our model that link the game's primitives to players' actions. We use one (Bayes Nash equilibrium) to describe the data-generating process and use the other (Bayes Correlated equilibrium) to describe the implications for observables while staying agnostic about the underlying information structure. The two concepts will be linked through the ordering of information structures introduced earlier.

The first is the standard concept of Bayes-Nash equilibrium.

Definition 6 (Bayes Nash Equilibrium)**.** *Let $S \in \mathcal{S}^*$. A strategy profile $\sigma = \left(\sigma_1, \ldots, \sigma_{|N|}\right)$ is a Bayes Nash Equilibrium (BNE) of the game $\Gamma^x(\theta, S)$ if for every $i \in N$, $\varepsilon_i \in \mathcal{E}_i$ and $\tau_i \in \mathcal{T}_i^x$ we have that, whenever for some $y_i \in Y_i$ the corresponding $\sigma_i\left(y_i \mid \varepsilon_i, t_i\right) > 0$, then :*

$$E_{\sigma_{-i}}\left[\pi_i\left(y_i, y_{-i}, \varepsilon_i; x, \theta_\pi\right) \mid \varepsilon_i, \tau_i\right] \geq E_{\sigma_{-i}}\left[\pi_i\left(y_i', y_{-i}, \varepsilon_i; x, \theta_\pi\right) \mid \varepsilon_i, \tau_i\right], \quad \forall\, y_i' \in Y_i,$$

where the expectation of y_{-i} is taken with respect to the distribution of equilibrium play $\sigma_{-i}\left(y_{-i} \mid \varepsilon_j, \tau_j\right) = \prod_{j \neq i} \sigma_j\left(y_j \mid \varepsilon_j, \tau_j\right)$.

Let the set $BNE^x(\theta, S)$ be the set of Bayes-Nash equilibria for game $\Gamma^x(\theta, S)$. The following set collects all conditional outcome distributions compatible with some BNE in the game.

Definition 7 (BNE Predictions)**.** *For a game $\Gamma^x(\theta, S)$, the set of* BNE predictions *is*

$$Q_{\theta,S}^{BNE}(x) = \left\{q(\cdot|x) \in \Delta^{|Y|} : q\left(y|x\right) = E[\sigma(y|\varepsilon, \tau)|x], \sigma \in BNE^x(\theta, S)\right\},$$

where expectation $E[\cdot|x]$ is with respect to the conditional distribution of (ε, τ) determined by the signal distribution $P_{\tau|\varepsilon}^x$ and prior $F(\cdot; \theta_\varepsilon)$.

We assume that players play a BNE under an unknown information structure S. Precisely characterizing the BNE predictions is not straightforward for two reasons. First, one needs to model the underlying information structure S. Furthermore, calculating $Q_{\theta,S}^{BNE}(x)$ requires finding all fixed points of the best-response conditions.

Although specifying S is hard, the researcher may suspect that it is at least as informative as some *baseline information structure* S^r. Is it then possible to link the ordering of information structures to predictions in a way that enables researchers to test this hypothesis? The answer is affirmative. For this, we employ a solution concept whose predictions respect the ordering of information structures. This alternative solution concept is called *Bayes correlated equilibrium* (BCE).[2] In what follows, we let $\mathcal{P}_W$ denote the space of probability distributions defined on the underlying space W. For example, $\mathcal{P}_{Y \times \mathcal{E} \times \mathcal{T}}$ represents the set of joint distributions of (y, ϵ, τ).

Definition 8 (Bayes Correlated Equilibrium)**.** *A Bayes Correlated Equilibrium $\nu \in \mathcal{P}_{Y \times \mathcal{E} \times \mathcal{T}}$ for the game $\Gamma^x(\theta, S)$ is a probability measure ν over action profiles, payoff types, and signals, which is*

1. Consistent with the prior*;*

$$\int_A \int_Y \nu^x(dy, d\varepsilon, dt) = \int_A P_{\tau|\varepsilon}^x(dt|\varepsilon) F(d\varepsilon; \theta_\varepsilon),$$

for any measurable $A \subset \mathcal{E} \times \mathcal{T}$; and

[2] This definition is slightly different than the one in Bergemann and Morris (2016). They define a decision rule to be a mapping

$$\sigma : \mathcal{T} \times \mathcal{E} \to \Delta(Y)$$

So a decision rule is a kernel that describes the probability of every action profile for a given vector of payoff types/signals (ε, τ). Hence, in Bergemann and Morris (2016), a BCE is a decision rule that satisfies obedience. To define obedience, they combine the decision rule with the common prior defined on payoff types ε and the prior distribution over signals τ. In their definition, Bergemann and Morris (2016) do not need to explicitly impose consistency as an equilibrium requirement (it is implicit in their setting). Instead, we define a BCE to be a distribution over actions, payoff types and types, and require consistency to hold.

2. Incentive compatible*; for all* $i, \varepsilon_i, \tau_i, y_i$ *such that* $\nu^x(y_i \mid \varepsilon_i, \tau_i) > 0$,

$$E_\nu\left[\pi_i\left(y_i, y_{-i}, \varepsilon_i; x, \theta_\pi\right) \mid y_i, \varepsilon_i, \tau_i\right] \geq E_\nu\left[\pi_i\left(y_i', y_{-i}, \varepsilon_i; x, \theta_\pi\right) \mid y_i, \varepsilon_i, \tau_i\right], \qquad \forall y_i' \in Y_i,$$

where the expectation operator $E_\nu[\cdot \mid y_i, \varepsilon_i, \tau_i]$ *is taken with respect to the conditional equilibrium distribution* $\nu(y_{-i}, \varepsilon_{-i}, \tau_{-i} \mid y_i, \varepsilon_i, \tau_i)$.

Here is how we interpret the BCE. First, the individuals play a BNE under unknown information structure S. From the analyst's point of view, their behavior is consistent with the following description.

1. There's a baseline information structure S^r. The players may know more than S^r;
2. A *mediator* observes $\varepsilon \sim F(\cdot; \theta_\pi)$ and $\tau \sim P^x_{\tau|\varepsilon}$ under S^r;
3. The mediator draws $y \sim \nu(y|\tau, \epsilon)$ and privately tells each i to play y_i.
4. The players *obey* the mediator's recommendation.

This view is convenient because we do not need to know the precise form of S as long as $S \succeq S^r$.

Let $BCE^x(\theta, S) \subset \mathcal{P}_{Y,\mathcal{E},\mathcal{T}}$ denote the set of all BCE distributions for the game $\Gamma^x(\theta, S)$. A BCE distribution is a complex object describing the joint distribution of actions and payoff types. For any such distribution ν we define a compatible *BCE prediction* $Q^{BCE}_{\theta,S}(x)$. This set collects the distributions of the observables that are compatible with BCE (Magnolfi and Roncoroni, 2022). Formally:

Definition 9 (BCE Predictions)**.** *For a game* $\Gamma^x(\theta, S)$*, the set of* BCE predictions *is:*

$$Q^{BCE}_{\theta,S}(x) = \left\{ q \in \Delta^{|Y|} : q(y) = \int_{\mathcal{E}\times\mathcal{T}} \nu\left(y, d\varepsilon, d\tau\right), \nu \in BCE^x(\theta, S) \right\}.$$

An advantage of working with the BCE prediction is that $Q^{BCE}_{\theta,S}(x)$ is a compact convex set because the incentive compatibility and consistency with prior impose linear restrictions on ν. We will use this feature to construct a computationally tractable test statistic. The robust prediction result of Bergemann and Morris (2016), which we state here for ease of reference, links the BCE and BNE predictions of the game as follows:

Lemma 2. *For all* $\theta \in \Theta$ *and* $x \in X$,

1. *If* $q \in Q^{BCE}_{\theta,S^r}(x)$, *then* $q \in Q^{BNE}_{\theta,S}(x)$ *for some* $S \in \mathcal{S}(S^r)$.
2. *Conversely, for all* $S \in \mathcal{S}(S^r)$, $Q^{BNE}_{\theta,S}(x) \subseteq Q^{BCE}_{\theta,S^r}(x)$.

Having discussed the main features of BCE, we can introduce a key property of the ordering induced by individual sufficiency, which has implications for the observables:

Theorem 1. $S^1(x)$ *is individually sufficient for* $S^2(x)$ *if and only if* $Q^{BCE}_{\theta,S^1}(x) \subseteq Q^{BCE}_{\theta,S^2}(x)$.

This theorem is an extension of Theorem 2 in Bergemann and Morris (2016), and spells out the consequences of individual sufficiency for players' behavior: if an information structure $S^1(x)$ is more informative (in the sense of individual sufficiency) of an information structure $S^2(x)$, the set of predictions corresponding to $S^1(x)$ is nested in the set of predictions implied by $S^2(x)$.

Theorem 1 allows a comparison of information structures based entirely on the sets of BCE predictions. This gives also a sense of when there is empirical content in testing an assumption about an extended information structure. If S^1 is (strictly) individually sufficient for S^2, then the set of predictions are a strict subset.[3] We discuss more in depth the implications of our results for testing in the next section.

3 Tests of Information Ordering

We consider the following question: can we test whether there is *more* information in the game than a certain information structure? Formally, we consider testing

$$H_0 : S \succeq S^r \quad \text{vs} \quad H_1 : S \not\succeq S^r. \tag{2}$$

Let us investigate the empirical contents of the null hypothesis. Note that (2) is a one-sided hypothesis regarding the unknown information structure. Working with the BCE prediction is attractive because it respects the ordering of the information structures via set inclusion relationships. This is not necessarily the case with the BNE prediction.

For example, one may test whether one of the players has privileged information in specific markets by choosing S^r properly, as in Example 3. An important aspect of our environment is its generality, which allows for the possibility of accommodating and testing information structures that vary across markets. As S specifies an information structure for every game x, the distribution of signals $P^x_{\tau|\varepsilon}$ may vary across markets x for a fixed S in the DGP. Consider now a restriction S^r, testing a statement such as $S \succeq S^r$: this restricts the heterogeneity of information structures across markets, as we are implicitly imposing $S(x) \succeq S^r$ for all x. In this case, the restriction in S^r is the same across all markets as in previous examples (e.g., complete information for every x). However, S^r may also be heterogeneous across markets as in Example 4 below.

Example 3: Consider the empirical setting of entry in airline markets similar to Ciliberto and Tamer (2009), where each market m is a city-pair, and airlines are players in a binary entry game. As in our application of Section 5, we focus on markets where at least one endpoint is a hub for a major airline, and we consider three separate players in each market: major airlines with a hub at least one endpoint, major airlines without hubs in the market, and low-cost airlines. Moreover, we include in the vector x_m airline-specific shifters of profitability such as measures of market presence and market size. Although $S(x)$ may depend on x_m, we specify a restriction on information S^r that does not depend on covariates. Specifically, for each market m, S^r prescribes that the hub airline player has complete information, whereas the other two players have incomplete information, receiving uninformative signals. In essence, this specification of S^r corresponds to the privileged information structure S_P defined in Section 2.1, and allows the researcher to test whether hubs confer informational advantages over competitors' cost and profit.

Example 4: Consider the empirical setting of entry in grocery markets similar to Magnolfi and Roncoroni (2016a), where supermarket firms are potential entrants in each geographic market m. Because firms may have political connections that affect their entry costs, the vector of market and firm-specific covariates includes $x^{Conn}_{i,m}$, a measure of the intensity of the connection. In addition, political connections may also

[3] By strict individually sufficient we mean that S^1 is individually sufficient for S^2, but S^2 is not individually sufficient for S^1.

affect the information structure of the game, e.g., by providing connected players with superior information about rivals' costs. More concretely, we could test a restriction on information $S^r(x)$ such that players whose $x_{i,m}^{Conn}$ exceeds a certain threshold have complete information, whereas their competitors with low $x_{i,m}^{Conn}$ have incomplete information. In this framework, the identity of connected players may vary across markets according to $x_{i,m}^{Conn}$, and more than one player may be connected in each market.

We note that our formulation of null and alternative in Equation 2 is well suited for some purposes (e.g., it makes sense to test the null hypothesis of a restriction on information if a researcher wants to analyze counterfactuals based on that restriction), but may not be appropriate for others. For instance, in Example 4 above, we would not impose the restriction of corruption or political connections as the null if the results are to be used to support an investigation into criminal behavior. Other formulations of the null are possible, and our framework may be extended to accommodate these.

3.1 Testable Implications, Test Statistics, and Asymptotic Properties

We make the following assumptions on the parameter space and data generating process.

Assumption 1. *(i) X is a finite set; (ii) $\Theta \subset \mathbb{R}^{d_\theta}$ is a compact set; (iii) $\mathcal{S}^*$ is equipped with the partial order $\succeq$.*

Assumption 1 sets regularity conditions on the parameter space and the support of covariates. Compactness of Θ is standard. The parameter space $\mathcal{S}^*$ for the information structure is partially ordered by the individual sufficiency ordering $\succeq$. For simplicity, we assume that the covariates are discrete (or discretized). [4] The next assumption states, in each market, the data are generated from a Bayes-Nash equilibrium under an unknown information structure S.

Assumption 2. *The conditional choice probability $P_{y|x}$ satisfies*

$$P_{y|x} \in Q_{\theta,S}^{BNE}(x), \tag{3}$$

for some $S \in \mathcal{S}^*$ and $\theta \in \Theta$.

Let P_x be the marginal law of x and let P be the joint law of (y, x). For each x, we identify $P_{y|x}$ with a vector in the $|Y| - 1$-dimensional simplex. We assume a sample is drawn from P independently across markets.

Assumption 3. *$(y^n, x^n) = (y_\ell, x_\ell)_{\ell=1}^n$ is a random sample from the law $P \in \mathcal{P}_{Y\times X}$.*

A common empirical practice is to specify the true information structure S and estimate θ using the restriction $P_{y|x} \in Q_{\theta,S}^{BNE}(x)$.[5] Instead, we specify the baseline information structure $S^r \in \mathcal{S}^*$ and test (2). If

[4] Allowing continuous covariates yields a continuum of conditional moment restrictions (see e.g. Chernozhukov et al., 2013). We leave this extension elsewhere to keep our tight focus on testing information structures.

[5] The existing work that employ the complete information structure as S includes Bjorn and Vuong (1984), Bresnahan and Reiss (1991), Berry (1992), Bajari et al. (2010), Ciliberto and Tamer (2009), and Ciliberto et al. (2018). Another commonly used specification is the incomplete information structure, in which each player only knows their own payoff state. A partial list of articles that employ this structure includes Seim (2006), Sweeting (2009), Aradillas-Lopez (2010), Bajari et al. (2007), and de Paula and Tang (2012).

H_0 is true, the the following testable implication holds by the assumptions above, Lemma 2, and Theorem 1:

$$P_{y|x} \in Q^{BNE}_{\theta,S}(x) \subseteq Q^{BCE}_{\theta,S^r}(x),\ \forall x \in X \tag{4}$$

We emphasize the tractability of the implied restriction: $P_{y|x} \in Q^{BCE}_{\theta,S^r}(x)$. Recall that $Q^{BNE}_{\theta,S}(x)$ is a mixture over the BNEs. It is not straightforward to see if the unknown information structure S is more informative than S^r by directly investigating $Q^{BNE}_{\theta,S}(x)$. Eq. (4) shows it suffices to check whether the conditional choice probability $P_{y|x}$ belongs to $Q^{BCE}_{\theta,S^r}(x)$. The conditional law $P_{y|x}$ can be recovered from the sample, and checking whether it belongs to the BCE prediction is computationally tractable.

For any $b \in \mathbb{R}^{|Y|}$ and a closed convex set A, let $h(b, A) = \sup_{a\in A} b^\top a$ be the *support function* of A. We may restate (4) as follows:

$$P_{y|x} \in Q^{BCE}_{\theta,S^r}(x) \quad \Leftrightarrow \quad b^\top P_{y|x} \le h(b, Q^{BCE}_{\theta,S^r}(x)),\ \forall b \in \mathbb{B}_x \text{ and } \forall x \in X, \tag{5}$$

where $\mathbb{B}_x = \{b \in \mathbb{R}^{|Y|} : \|b\|_{W_x} \le 1\}$ is the unit ball for a norm $\|\cdot\|_{W_x}$ defined below. Observe that we translated the hypothesis on the ordering of information structures into the ordering of functions: $(b, x) \mapsto b^\top P_{y|x}$ and $(b, x) \mapsto h(b, Q^{BCE}_{\theta,S^r}(x))$. Using this, one can implement a test by comparing linear combinations of $P_{y|x}$ and the support function of the BCE prediction. For each θ, let

$$T(\theta) \equiv \sup_{x\in X} \sup_{b\in\mathbb{B}_x} \{b^\top P_{y|x} - h(b, Q^{BCE}_{\theta,S^r}(x))\}. \tag{6}$$

As $\mathbb{B}_x$ contains the origin, $T(\theta)$ is always non-negative. Moreover, when the null hypothesis is true, it takes value zero for all θ.[6]

Our test is based on a sample analog of $T(\theta)$. Statistics based on support functions are used in various partially identified models (Beresteanu and Molinari, 2008; Beresteanu et al., 2011; Bontemps et al., 2012; Kaido and Santos, 2014). The construction of our test statistic is most closely related to that of Beresteanu et al. (2011), but it has a novel feature. By introducing the weighted norm to define $\mathbb{B}_x$, one can studentize the moments used to construct the test statistic. Let $\hat{P}_{n,x} \in \mathbb{R}^{|Y|}$ denote the empirical conditional choice probability whose j-th component is

$$\hat{P}^{(j)}_{n,x} = \frac{1}{n_x} \sum_{\ell=1}^{n} 1\{y_\ell = y^{(j)}, x_\ell = x\}, \tag{7}$$

where $n_x = n^{-1} \sum_{\ell=1}^{n} 1\{x_\ell = x\}$. Let $W_{n,x} \equiv n\widehat{var}(\hat{P}_{n,x})$ be the sample covariance matrix of $\hat{P}_{n,x}$, and assume $W_{n,x}$ is positive definite. Let

$$\mathbb{B}_{n,x} \equiv \{b \in \mathbb{R}^{|Y|} : \|b\|_{W_{n,x}} \le 1\},\ \|b\|_{W_{n,x}} \equiv (b^\top W_{n,x} b)^{1/2}. \tag{8}$$

Define

$$T_n(\theta) \equiv \sup_{x\in X} \sup_{b\in\mathbb{B}_{n,x}} \sqrt{n}\{b^\top \hat{P}_{n,x} - h(b, Q^{BCE}_{\theta,S^r}(x))\}. \tag{9}$$

[6]This is because, if $b^\top P_{y|x} - h(b, Q^{BCE}_{\theta,S^r}(x)) \le 0$ for all $b \in \mathbb{B}_x$, the maximum value is achieved by setting $b = 0$.

This is the sample counterpart of $T(\theta)$. This statistic can also be expressed as follows (see Lemma 3 in Appendix):

$$\sup_{b\in\mathbb{B}_{n,x}}\sqrt{n}\{b^{\top}\hat{P}_{n,x}-h(b,Q^{BCE}_{\theta,S^r}(x))\}=\sup_{b\in\mathbb{R}^{|Y|}\setminus\{0\}}\sqrt{n}\Big\{\frac{b^{\top}\hat{P}_{n,x}-h(b,Q^{BCE}_{\theta,S^r}(x))}{\sqrt{nb^{\top}W_{n,x}b}}\Big\}_{+}. \tag{10}$$

Hence, for each b, the quantity $\sqrt{n}\{b^{\top}\hat{P}_{n,x}-h(b,Q^{BCE}_{\theta,S^r}(x))$ is divided by its standard error. This is desirable because T_n constructed this way is less sensitive to imprecisely estimated moments than a statistic without any studentization. It can be shown that $T_n(\theta)=0$ whenever $\hat{P}_{n,x}\in Q^{BCE}_{\theta,S^r}(x)$ for all x. When $\hat{P}_{n,x}$ is outside the BCE prediction, the statistic measures the maximum deviation of the empirical distribution from the BCE prediction for each value of θ. Using techniques in Magnolfi and Roncoroni (2016b), this statistic can be calculated by solving a convex program. We provide details on computational aspects in Section 3.2.

The behavior of the test statistic T_n depends on where $P_{y|x}$ is located relative to the BCE prediction. It is desirable to design a test that remains valid over a large class of DGPs. We therefore combine bootstrap with a moment selection procedure. This approach builds on the developments in the moment inequalities literature (e.g., Andrews and Soares (2010), Andrews and Barwick (2012), Bugni et al. (2015)).

Let us rewrite T_n as

$$T_n(\theta)=\sup_{x\in X}\sup_{b\in\mathbb{B}_{n,x}}\{\mathbb{G}_n(b,x)+\eta_\theta(b,x)\}, \tag{11}$$

where $\mathbb{G}_n(b,x)=\sqrt{n}b^{\top}(\hat{P}_{n,x}-P_{y|x})$, and $\eta_\theta(b,x)\equiv\sqrt{n}(b^{\top}P_{y|x}-h(b,Q^{BCE}_{\theta,S^r}(x)))$. The values of (b,x) that are relevant for the supremum in (11) are the ones such that $\eta_\theta(b,x)$ is close to zero. The challenge is that $\eta_\theta(b,x)$ cannot be consistently estimated uniformly over DGPs. We, therefore, proceed as follows.

Let $\hat{\eta}_{n,\theta}(b,x)\equiv\sqrt{n}(b^{\top}\hat{P}_{n,x}-h(b,Q^{BCE}_{\theta,S^r}(x)))$ and let $\{\tau_n\}$ be a slowly diverging sequence such as $\tau_n=\sqrt{\ln n}$. Define

$$\Psi_{n,\theta}\equiv\{(b,x):\hat{\eta}_{n,\theta}(b,x)\geq-\tau_n, b\in\mathbb{B}_{n,x}, x\in X\}\quad. \tag{12}$$

This is the set of (b,x) pairs for which the sample moment restrictions are nearly binding. Here, we follow the insights of generalized moment selection procedures (Andrews and Soares, 2010; Chernozhukov et al., 2013) to conservatively approximate the set of binding moment restrictions. Let (y^{*n},x^{*n}) be a bootstrap sample drawn from the empirical distribution $\hat{P}_n$.[7] Let $\hat{P}^*_{n,x}\in\mathbb{R}^{|Y|}$ be a vector of empirical conditional frequencies of outcomes in a bootstrap sample given $x^*=x$. Let $\mathbb{G}^*_n\equiv\sqrt{n}b^{\top}(\hat{P}^*_{n,x}-\hat{P}_{n,x})$. A bootstrap analog of T_n is defined by

$$T^*_n(\theta)\equiv\sup_{(b,x)\in\Psi_{n,\theta}}\{\mathbb{G}^*_n(b,x)\}. \tag{13}$$

Recall that $T(\theta)$ in (6) takes the value zero for any $\theta\in\Theta$ under the null hypothesis. A valid level-α test is the one that tests $T(\theta)=0$ individually at level α and rejects H_0 only if the tests reject $T(\theta)=0$ for all $\theta\in\Theta$. Such a test controls size but can be conservative. Nonetheless, it can still be attractive for several practical reasons. First, it yields a confidence set for the structural parameter θ when H_0 is not rejected.

[7] The empirical distribution is $\hat{P}_n=\frac{1}{n}\sum_{\ell=1}^{n}\delta_{(y_i,x_i)}$, where $\delta_{(y,x)}$ is a point mass at (y,x).

Second, the computational costs of this test can be mitigated by combining it with a Bayesian optimization algorithm, which we elaborate in the next section.[8]

For each $\theta \in \Theta$, let the bootstrap p-value be

$$p_n(\theta) \equiv P^*(T_n^*(\theta) > T_n(\theta)|y^n, x^n), \tag{14}$$

where P^* is the distribution of T_n^* conditional on the sample (y^n, x^n). We reject H_0 if

$$p_n(\theta) \leq \alpha \quad \text{for all } \theta \in \Theta. \tag{15}$$

Otherwise, there exists a parameter value $\theta \in \Theta$ for which an information structure $S \succeq S^r$ is consistent with data. We may then collect such parameter values to define a confidence set:

$$CS_n \equiv \{\theta \in \Theta : p_n(\theta) > \alpha\}. \tag{16}$$

This set covers θ with probability $1 - \alpha$ asymptotically.

Theorem 2 below ensures that the test is asymptotically valid over a large class of data-generating processes. For any M-by-M matrix A, let $\|A\|_{op} = \inf\{c \geq 0 : \|Az\| \leq c\|z\|, \ \forall z \in \mathbb{R}^M\}$ be its operator norm. We impose the following regularity conditions.

Assumption 4. *There exists $W_x \in \mathbb{R}^{|Y| \times |Y|}$ with $x \in X$ such that, for any $\epsilon > 0$,*

$$P(\sup_{x \in X} \|W_{n,x} - W_x\|_{op} > \epsilon) \leq \epsilon, \ \forall n \geq N_\epsilon,$$

for some $N_\epsilon \in \mathbb{N}$; (ii) $\underline{\kappa} < \underline{\lambda}(W_x)$ and $\overline{\lambda}(W_x) < \overline{\kappa}$ for uniform constants $0 < \underline{\kappa} \leq \overline{\kappa} < \infty$ for all $x \in X$, where $\underline{\lambda}(\cdot)$ and $\overline{\lambda}(\cdot)$ are the smallest and largest eigenvalues, respectively.

Assumption 5. *There is $\underline{\zeta} > 0$ such that $P_{y|x}(y_\ell = y|x) \geq \underline{\zeta}$ and $P(x_\ell = x) \geq \underline{\zeta}$ for all $y \in Y$ and $x \in X$.*

Assumption 4 states the sample weighting matrix converges uniformly to its limiting counterpart W_x. We also assume W_x's eigenvalues are uniformly bounded away from 0 and from above. Assumption 5 requires y_i's conditional probability and x_i's probability mass function are uniformly bounded away from 0.

For each $\theta \in \Theta$, we define our null model $\mathcal{P}_\theta$ as follows:

$$\mathcal{P}_\theta \equiv \Big\{P \in \mathcal{P}_{Y \times X} : \text{ Assumptions 2-5 hold with respect to } P\Big\}. \tag{17}$$

Then, by Assumption 2, (y^n, x^n) is generated from $P \in \mathcal{P}_\theta \subseteq \mathcal{P}_{Y \times X}$. Let $\mathcal{P} \equiv \{P : P \in \mathcal{P}_\theta, \ \theta \in \Theta\}$ be the set of all distributions compatible with our restrictions.

Theorem 2. *Suppose Assumptions 1-5 hold. Suppose $\tau_n > 0$ for all n, $\tau_n \to \infty$ and $\tau_n = o(n^{1/2})$. Let $\alpha \in (0, 1/2)$. Let $\phi(y^n, x^n) = 1\big\{\sup_{\theta \in \Theta} p_n(\theta) \leq \alpha\big\}$. Then,*

$$\limsup_{n \to \infty} \sup_{P \in \mathcal{P}} E_P[\phi] \leq \alpha. \tag{18}$$

[8]An alternative approach would be to use $\inf_{\theta \in \Theta} T_n(\theta)$ as a test statistic for H_0. This approach requires resampling a suitable analog of $\inf_{\theta \in \Theta} T_n(\theta)$, which can be computationally demanding although it may lead to a potential gain in power. Also, this approach is not suitable to construct a confidence region.

3.2 Computational Aspects

We discuss ways to simplify the computation of the statistic, implementation of the test, and construction of confidence intervals. Let

$$\begin{aligned} V_{n,x}(\theta) &\equiv \sup_{b\in\mathbb{B}_{n,x}} \sqrt{n}\{b^\top \hat{P}_{n,x} - h(b, Q^{BCE}_{\theta,S^r}(x))\} \\ &= \sup_{b\in\mathbb{B}_{n,x}} \inf_{q\in Q^{BCE}_{\theta,S^r}(x)} \left[b^\top \hat{P}_{n,x} - b^\top q\right], \quad (P0), \end{aligned}$$

and note that $T_n(\theta) = \sup_{x\in X} V_{n,x}(\theta)$. Following Magnolfi and Roncoroni (2016b), we may recast $(P0)$ as a convex quadratic program:

$$\begin{aligned} V_{n,x}(\theta) = \max_{lb\le w\le ub} \quad & -\gamma^\top w \\ s.t. \quad & w^\top \Gamma_1 w \le 1 \\ & \Gamma_2 w = 0_{|Y|} \\ & \Gamma_3 w \le 0_{d_\nu}, \end{aligned} \tag{19}$$

for some vector γ and matrices Γ_l for $l = 1, 2, 3$; we provide details on how to construct these objects in Appendix A.1. The vector $w = (b^\top, \lambda_{eq}^\top, \lambda_{ineq}^\top)^\top$ stacks $b \in \mathbb{R}^{|Y|-1}$ together with Lagrange multipliers associated with the constraints in the original problem. This reformulation simplifies the computation of the test statistic and, thus, the p-value function. Specifically, we solve the convex program for each (θ, x) and optimize the p-value function only once. This saves the number of times we need to solve a global non-convex optimization problem.[9]

Second, testing H_0 can be done by combining the p-value process $\theta \mapsto p_n(\theta)$ with a global optimization algorithm. For testing H_0, it suffices to check whether $\sup_\theta p_n(\theta)$ is below or above α. While p_n does not have a closed form in general, algorithms for optimizing such "black-box" functions are available. Among them, the *response surface method* is known to be able to globally optimize such functions without evaluating p_n many times (Jones et al., 1998).

Finally, a version of the response surface algorithm can also be applied to the computation of confidence intervals (i.e., coordinate projections of CS_n) for particular components of θ. Instead of using a grid, one can obtain the end points of a confidence interval by solving

$$\max / \min_{\theta\in\Theta} e_j^\top \theta \tag{20}$$

$$s.t.\ p_n(\theta) \ge \alpha, \tag{21}$$

where e_j is a vector of 0's whose j-th component is 1. This optimization problem has a black-box function in its constraint. A response surface algorithm for solving such constrained optimization problems is developed in Kaido et al. (2017).

[9] In contrast, directly applying a bootstrap procedure to the statistic $\inf_{\theta\in\Theta} T_n(\theta)$ is computationally demanding. It requires finding the global minimum of a non-linear (and typically non-convex) function $T_n(\theta)$ for a large number of times (essentially the number of bootstrap replications).

3.3 Multiple Testing

When the information structure is of independent economic interest, the analyst may consider testing a sequence of hypotheses to refine her understanding of the game's information structure. To motivate this, let us revisit Example 3. Recall that $H_0 : S \succeq S^r$ in this example is formed with $S^r = S_P$ to investigate whether the hub airline player has informational advantage over the others. Yet, one may argue that, even if H_0 is not rejected, it does not definitively conclude that the hub player has informational privilege, as the scenario where all the players participate in an entry game of complete information remains plausible. This is because $S_C \succeq S_P$. Now, suppose the analyst opts to test two nulls of the form: $H_{0,j} : S \succeq S_j^r$ for $j = 1, 2$ where $S_1^r = S_C$ and $S_2^r = S_P$. If $H_{0,2}$ is not rejected (as before) while $H_{0,1}$ is, then such test results may offer stronger evidence of the hub's privileged information. To test each $H_{0,j}$, one can directly employ a p-value calculated in Section 3.1. Given the test of multiple hypotheses, a remaining crucial task is to control the accumulated error of false rejections, namely, the family-wise error rate (FWER).

Consider a sequence of null hypotheses, $H^1, ..., H^J$ using the abbreviation $H^j \equiv H_0^j$, where

$$H^j : S \succeq S_j^r. \tag{22}$$

Motivated by the example above, we assume the following:

Assumption 6. $S_1^r \succeq \cdots \succeq S_J^r$.

Assumption 6 establishes a logical relationship among the nulls; for instance, if $H_{0,1}$ holds true, $H_{0,2}$ is automatically true. As shown below, this assumption provides additional structure in multiple testing, thereby improving the performance of the test. Instead of sequentially testing $H_{0,1}$ and $H_{0,2}$, one may wish to test a hypothesis of the form $S_1^r \succeq S \succeq S_2^r$. Although this form of hypothesis cannot be accommodated within our framework, we do not necessarily view this as a limitation of our framework. Rather, we view this as reflecting the robustness property of the framework, which can be seen, for instance, from Lemma 2.

To control the FWER, we propose the following step-wise procedure. This procedure is less conservative than a classic Bonferroni procedure that uses a common cutoff to be compared to p-values for all nulls. The procedure adapts Holm (1979)'s procedure to incorporate logical dependence (Assumption 6)—thus is less conservative than Holm's—and to accommodate the specification test with a partially identified model.

Similar to Section 3.1, $H^j : S \succeq S_J^r$ only if $H^j : T_{n,j}(\theta) = 0,\ \forall\theta$ where $T_{n,j}$ is the test statistic analogous to (11) for H^j. Let $p_j \equiv p_{n,j} \equiv \sup_{\theta\in\Theta} p_{n,j}(\theta)$ where $p_{n,j}(\theta)$ is the p-value analogous to (14) that corresponds to H^j. Let $p_{(1)} \leq \cdots \leq p_{(J)}$ be the ordered p_j's and $H^{(1)}, ..., H^{(J)}$ be the corresponding null hypotheses. Also let $c_1 \leq \cdots \leq c_J$ be an increasing sequence of constants. We propose to choose $c_j = \alpha/(j^* - j + 1)$ where j^* is such that $p_{(j)} = p_{j^*}$ and α is the nominal level.

Algorithm 1 (Modified Holm procedure)**.** *Proceed as follows.*

Step 1: If $p_{(1)} > c_1$, then stop (and reject no nulls). Otherwise, reject $H^{(1)}$ and go to Step 2.

$\vdots$

Step j: If $p_{(j)} > c_j$, then stop. Otherwise, reject $H^{(j)}$ and go to Step $j + 1$.

$\vdots$

Step J: If $p_{(J)} > c_J$, then stop. Otherwise, reject $H^{(J)}$ and stop.

The original Holm's procedure proposes $\tilde{c}_j = \alpha/(J-j+1)\ \forall j$. Since $c_j \geq \tilde{c}_j$, the test based on Algorithm 1 can be less conservative and more powerful than Holm's test. Under Assumption 6, note that even if H_j is rejected, it may not be a good idea to reject H_{j-k} ($\forall k$ such that $j-k \geq 1$) because H_j may have been falsely rejected. Instead, we incorporate the logical relationship in the algorithm by appropriately choosing c_j.[10]

Let $\phi^{MH}(y^n, x^n) = 1\{\text{Algorithm 1 rejects some nulls}\}$. The following theorem shows that the proposed test controls size.

Theorem 3. *Suppose Assumptions 1-6 hold. The modified Holm's procedure satisfies*

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}} E_P[\phi^{MH}] \leq \alpha.$$

Although the theory allows for a general J, we note that the $J = 2$ case can already yield more convincing and interpretable conclusions from test outcomes compared to the $J = 1$ case, as the example above suggests.

Continuing the example above, one may wish to alternatively set $S_1^r = \tilde{S}_P$ and $S_2^r = S_P$, where S_P is the information structure with privilege to the hub player and $\tilde{S}_P$ is that with privilege to the other players. Again, if $H_{0,1}$ is rejected while $H_{0,2}$ is not, one can at least confidently rule out the possibility of not rejecting $H_{0,2}$ because of the complete information of all the players. However, these nulls do not satisfy Assumption 6, and thus the test cannot exploit the logical relationship. We can still apply the original Holm's procedure, modified for the specification test with a partially identified model.

4 Monte Carlo Experiments

4.1 Simulation Design

To illustrate our method, we construct a class of information structures where players receive a private signal in addition to knowing their own payoff types. In the limit, the private signal is perfectly informative about a portion of the opponent's payoff type, making it public information as in S_{Pub} in Example 2. This limit information structure is adopted by Aguirregabiria and Mira (2010) in the context of dynamic entry games.

To fix ideas, consider the two-player entry game of Example 1, where firm i's profits are zero when not entering, and $\pi_i(y, \varepsilon_i; x, \theta_\pi) = x\beta + \Delta y_{-i} + \varepsilon_i$ upon entry, where x is a random variable supported on $\{-M, M\}$. Suppose that payoff type for player i is $\varepsilon_i = \nu_i + \epsilon_i$. Let ν_i take values in $\{-\eta, \eta\}$ with $P(\nu_i = \eta) = p \in [0, 1]$, where η is a positive scalar. The error component ν_i is i.i.d. across players. The other component ϵ_i is a standard normal random variable, which is also independent across i. The structure of the game is common knowledge among the players.

Consider an information structure where, in addition to knowing their own payoff type, each player gets an informative signal t_i, with support $\{\eta, -\eta\}$ which relays information on the realization of ν_{-i}. In particular, the signal reveals ν_{-i} with probability q, or $P(t_i = \nu \mid \nu_{-i} = \nu) = q$ for $\nu \in \{\eta, -\eta\}$. Thus, upon

[10]We may be able to further improve upon our procedure by incorporating the knowledge of the joint distribution of the p-values (e.g., Romano and Wolf, 2005), which is beyond the scope of the paper.

observing $t_i = \eta$, player i will conclude that $\nu_{-i} = \eta$ with probability:

$$\begin{aligned} P(\nu_{-i} = \eta \mid t_i = \eta) &= \frac{P(\nu_{-i} = \eta)P(t_i = \eta \mid \nu_{-i} = \eta)}{P(t_i = \eta)} \\ &= \frac{pq}{pq + (1-p)(1-q)} \equiv \rho_\eta(p, q). \end{aligned}$$

Call this information structure S_q. Note that, as $q \to \frac{1}{2}, \rho_\eta(p,q) \to p$. Thus, as q approaches $1/2$, S_q approaches S_I, the incomplete information structure, under which each player only knows their own type. Similarly,

$$\begin{aligned} P(\nu_{-i} = -\eta \mid t_i = -\eta) &= \frac{P(\nu_{-i} = -\eta)P(t_i = -\eta \mid \nu_{-i} = -\eta)}{P(t_i = -\eta)} \\ &= \frac{(1-p)q}{(1-p)q + p(1-q)} \equiv \rho_{-\eta}(p, q). \end{aligned}$$

For this information structure, Bayes Nash equilibria are characterized by the threshold strategy:

$$y_i = 1\big\{\varepsilon_i \geq \tau_i(\nu_i, t_i)\big\},\ i = 1, 2, \tag{23}$$

where $\tau_i(\nu_i, t_i)$ is defined in Appendix C. Let $\Psi(\tau) = P(\varepsilon_i \geq \tau)$. The equilibrium probability of entry for one player conditional on a state ν_i is:

$$\begin{aligned} P_{\nu_i=\eta,x} &= q\Psi(\tau(\eta,\eta)) + (1-q)\Psi(\tau(\eta,-\eta)) \\ P_{\nu_i=-\eta,x} &= q\Psi(\tau(-\eta,-\eta)) + (1-q)\Psi(\tau(-\eta,\eta)). \end{aligned}$$

We simulate a BNE outcome using the conditional probability

$$P_{y|x}(y|x) = \int \prod_{i=1}^{2} (P_{\nu_i=\eta,x})^{y_i}(1 - P_{\nu_i=\eta,x})^{(1-y_i)} dP(\nu), \quad y = (y_1, y_2)^\top \in Y. \tag{24}$$

The equilibrium conditional choice probability (CCP) depends on the informativeness of the signal measured by $q \in [0.5, 1]$. Figure 1 (top panel) shows how the equilibrium conditional choice probability changes with q. The entire CCP vector changes from the CCPs under low-quality signals (blue dots) to those under high-quality signals (yellow dots). We test the null hypothesis that $S \succeq S_{Pub}$. That is, S is at least as informative as the public information structure S_{Pub}. Under S_{Pub}, ν_i's are publicly known. The given information structure S_{Pub} corresponds to the setting with $q = 1$. Figure 1 also plots the BCE prediction (for a fixed θ). The BCE prediction contains the CCP with $q = 1$ as expected. It also contains some of the CCPs under relatively high-quality signals about ν_{-i}. Hence, we cannot expect any test (including ours) to have power against alternatives that are close in terms of the signal quality to the null hypothesis. However, we should expect that one can attain meaningful power when the equilibrium CCP is outside the BCE prediction under S^r. This happens when q sufficiently departs from 1 towards 0.5. Below, we examine the size and power of the test when data are generated from a BNE with q varying over $[0.5, 1]$.

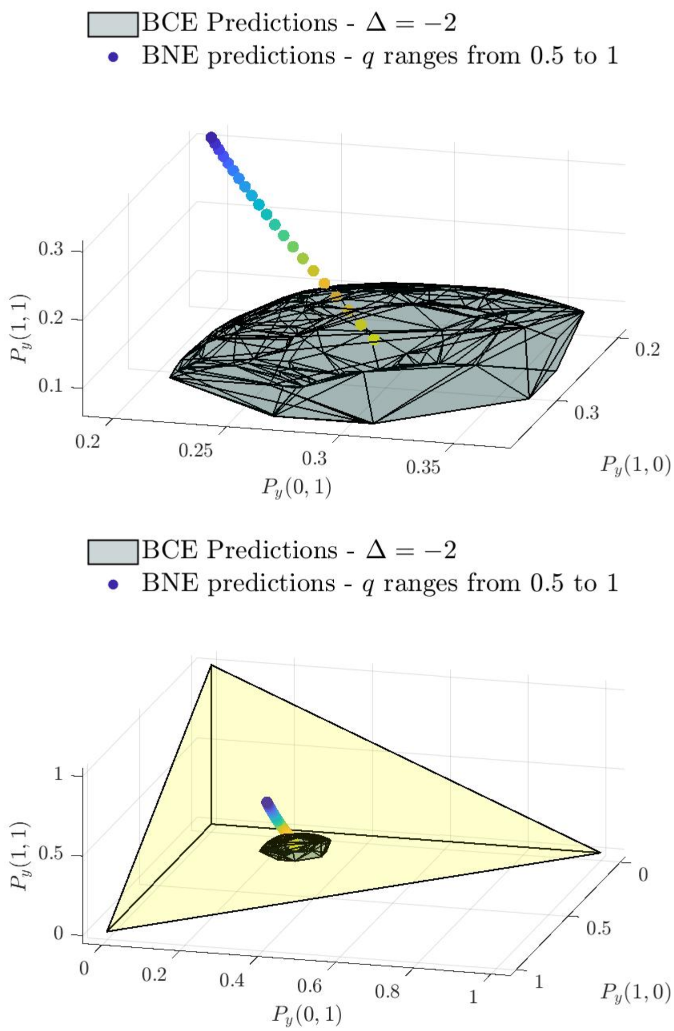

Figure 1: Equilibrium CCPs and the BCE Prediction

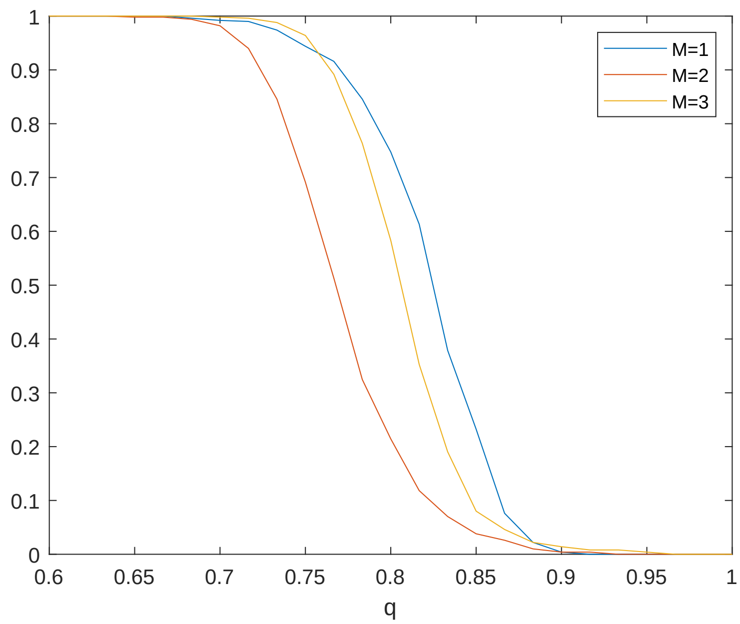

Figure 2: The Rejection Probability of the Test

4.2 Simulation Results

Figure 2 shows the power of the proposed test for a sample of size 1000 with $M = 1, 2$, and 3. The null hypothesis corresponds to $q = 1$. We set the significance level to 0.05. The test controls the size across all values of M, but it tends to be conservative at $q = 1$. This is expected. Recall that the conditional choice probability remained inside the BCE prediction for high values of q in Figure 1. Consistent with this observation, the rejection probability of the test starts increasing only after q becomes sufficiently small. The rejection probability then grows rapidly as q gets smaller, demonstrating that the test can detect local deviations from the boundary of the BCE prediction. Figure 2 also shows a nonmonotonic interplay of q and the support of the covariates. When q is sufficiently high or low (i.e., q above 0.9 or below 0.75), setting M to a large value ($M = 3$) makes the rejection probability of the test slightly higher than the power under the other designs. On the other hand, when q is in the range from roughly 0.775 to 0.875, the test achieves higher power with $M = 1$. Interestingly, the test has the lowest power for almost all values of q when $M = 2$.

5 Empirical Application: Information in Airline Markets (in progress)

The US airline industry quickly evolved towards an hub-and-spoke model after deregulation in the 1970s. Berry et al. (1996) find that hubs provide two distinct advantages. On the demand side, a hub airline can provide amenities to business travelers that generate market power with respect to these inelastic consumers. On the cost side, hubbing generates cost economies, especially on longer routes. In this application, we investigate another, separate potential advantage of hubbing: better information at the entry stage. When considering entry on potential new routes that originate in its hub, the hub airline may benefit from superior ability to forecast demand, and better understanding of costs - both for its own operations and for its rivals'. In turn, this informational advantage, added to the demand and cost advantages that affect profits upon entry, may make it particularly hard to compete with hub airlines.

6 Conclusion

The information available to players is a key primitive of any strategic environment. While it is hard to specify the information structure precisely, data can provide guidance on the baseline information structure that allows us to conduct robust policy evaluations and other counterfactual analyses. This paper develops a formal testing framework on the baseline information structure while treating it as a nonparametric object that varies with the game's observable characteristics.

References

Aguirregabiria, V. and P. Mira (2010): "Dynamic discrete choice structural models: A survey," *Journal of Econometrics*, 156, 38–67, structural Models of Optimization Behavior in Labor, Aging, and Health.

Andrews, D. W. (1994): "Chapter 37 Empirical process methods in econometrics," Elsevier, vol. 4 of *Handbook of Econometrics*, 2247 – 2294.

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

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

Aradillas-Lopez, A. (2010): "Semiparametric estimation of a simultaneous game with incomplete information," *Journal of Econometrics*, 157, 409–431.

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

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

Baltrunaite, A. (2020): "Political contributions and public procurement: evidence from Lithuania," *Journal of the European Economic Association*, 18, 541–582.

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

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

Bergemann, D., B. Brooks, and S. Morris (2017): "First-price auctions with general information structures: Implications for bidding and revenue," *Econometrica*, 85, 107–143.

BERGEMANN, D. AND S. MORRIS (2016): "Bayes correlated equilibrium and the comparison of information structures in games," *Theor Econ*, 11, 487–522.

BERRY, S., M. CARNALL, AND P. T. SPILLER (1996): "Airline Hubs: Costs, Markups and the Implications of Customer Heterogeneity," Working Paper 5561, National Bureau of Economic Research.

BERRY, S. T. (1992): "Estimation of a Model of Entry in the Airline Industry," *Econometrica: Journal of the Econometric Society*, 889–917.

BHATIA, R. (1997): *Matrix Analysis*, Springer New York.

BJORN, P. A. AND Q. H. VUONG (1984): "Simultaneous Equations Models for Dummy Endogenous Variables: A Game Theoretic Formulation with an Application to Labor Force Participation," California Institute of Technology, Social Science Working Paper.

BONTEMPS, C., T. MAGNAC, AND E. MAURIN (2012): "Set identified linear models," *Econometrica*, 80, 1129–1155.

BRESNAHAN, T. F. AND P. C. REISS (1991): "Empirical models of discrete games," *Journal of Econometrics*, 48, 57–81.

BROCK, W. A. AND S. N. DURLAUF (2001): "Discrete choice with social interactions," *The Review of Economic Studies*, 68, 235–260.

BUGNI, F. A., I. A. CANAY, AND X. SHI (2015): "Specification tests for partially identified models defined by moment inequalities," *Journal of Econometrics*, 185, 259–282.

CANAY, I. A. AND A. M. SHAIKH (2017): *Practical and Theoretical Advances in Inference for Partially Identified Models*, Cambridge University Press, 271–306, Econometric Society Monographs.

CHERNOZHUKOV, V., S. LEE, AND A. M. ROSEN (2013): "Intersection bounds: estimation and inference," *Econometrica*, 81, 667–737.

CILIBERTO, F., C. MURRY, AND E. TAMER (2018): "Market Structure and Competition in Airline Markets," *University of Virginia, Penn State University, Harvard University.*

CILIBERTO, F. AND E. TAMER (2009): "Market structure and multiple equilibria in airline markets," *Econometrica*, 77, 1791–1828.

DE PAULA, A. (2013): "Econometric Analysis of Games with Multiple Equilibria," *Annual Review of Economics*, 5, 107–131.

DE PAULA, Á. AND X. TANG (2012): "Inference of Signs of Interaction Effects in Simultaneous Games With Incomplete Information," *Econometrica*, 80, 143–172.

DICKSTEIN, M. J. AND E. MORALES (2018): "What do Exporters Know?*," *The Quarterly Journal of Economics*, 133, 1753–1801.

GALICHON, A. AND M. HENRY (2011): "Set identification in models with multiple equilibria," *The Review of Economic Studies*, 78, 1264–1298.

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

Gualdani, C. and S. Sinha (2023): "Identification in discrete choice models with imperfect information," ArXiv:1911.04529.

Holm, S. (1979): "A simple sequentially rejective multiple test procedure," *Scandinavian journal of statistics*, 65–70.

Jones, D. R., M. Schonlau, and W. J. Welch (1998): "Efficient global optimization of expensive black-box functions," *Journal of Global optimization*, 13, 455–492.

Jovanovic, B. (1989): "Observable implications of models with multiple equilibria," *Econometrica: Journal of the Econometric Society*, 1431–1437.

Kaido, H., F. Molinari, J. Stoye, and M. Thirkettle (2017): "Calibrated Projection in MATLAB: Users' Manual," *arXiv preprint arXiv:1710.09707*.

Kaido, H. and A. Santos (2014): "Asymptotically efficient estimation of models defined by convex moment inequalities," *Econometrica*, 82, 387–413.

Li, S., Y. Ogura, and V. Kreinovich (2002): *Limit Theorems and Apllications of Set-Valued and Fuzzy Set-Valued Random Variables*, Kluwer Academic Press.

Magnolfi, L. and C. Roncoroni (2016a): "Political connections and market structure," Tech. rep., Working paper.

——— (2016b): "Political Connections and Market Structure," .

——— (2022): "Estimation of Discrete Games with Weak Assumptions on Information," *The Review of Economic Studies*, 90, 2006–2041.

Mertens, J.-F. and S. Zamir (1985): "Formulation of Bayesian analysis for games with incomplete information," *International Journal of Game Theory*, 14, 1–29.

Molchanov, I. S. (2005): *Theory of random sets*, vol. 19, Springer.

Navarro, S. and Y. Takahashi (2012): "A Semiparametric Test of Agent's Information Sets for Games of Incomplete Information," Discussion Paper Series of SFB/TR 15 Governance and the Efficiency of Economic Systems 432, Free University of Berlin, Humboldt University of Berlin, University of Bonn, University of Mannheim, University of Munich.

Seim, K. (2006): "An empirical model of firm entry with endogenous product-type choices," .

Sweeting, A. (2009): "The strategic timing incentives of commercial radio stations: An empirical analysis using multiple equilibria," .

Syrgkanis, V., E. Tamer, and preprint arXiv:1710.03830, Ziani J (2017): "Inference on Auctions with Weak Assumptions on Information," *arXiv preprint arXiv:1710.03830*.

Tamer, E. (2003): "Incomplete simultaneous discrete response model with multiple equilibria," *The Review of Economic Studies*, 70, 147–165.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes*, Springer New York.

van Handel, R. (2018): "Probability in High Dimension," Princeton University.

Vershynin, R. (2018): *High-dimensional probability: An introduction with applications in data science*, vol. 47, Cambridge University Press.

A Computation

A.1 Computing $V_{n,x}(\theta)$

The object $V_{n,x}(\theta)$ is the value of the following maxmin program

$$V_{n,x}(\theta) = \max_{b \in \mathbb{B}_{n,x}} \min_{q \in Q^{BCE}_{\theta,S^r}(x)} \left[b^\top \hat{P}_{n,x} - b^\top q \right], \quad (P0)$$

which can be computed for every value of x and θ.

Step 1 - Discretization: To make $(P0)$ feasible we approximate the infinite dimensional object ν by discretizing the set $\mathcal{E}$. Let $\mathcal{E}^r \subset \mathcal{E}$ be the discretized set, with $|\mathcal{E}^r| = r$. We obtain $\mathcal{E}^r$ as the product space of $\mathcal{E}^r_i$, where every set $\mathcal{E}^r_i$ contains $r_i = \frac{r}{|N|}$ equally spaced quantiles of F_{ε_i}.[11] We also define $f^r(\cdot;\theta_\varepsilon)$ as the probability mass function over $\mathcal{E}^r$, where the mass of each $\varepsilon \in \mathcal{E}^r$ is generated by a Normal copula with correlation parameter $\rho = \theta_\varepsilon$. The program $(P0)$ is then approximated by the feasible program

$$\begin{array}{rcl}
\max\limits_{b \in \mathbb{R}^{|Y|}} \min\limits_{q \in \mathbb{R}^{|Y|}, \nu \in \mathbb{R}^{|Y| \times r}} & b^\top \left(\hat{P}_{n,x} - q \right) & (P1) \\
s.t. & b^\top W_{n,x} b - 1 & \leq 0 \\
\forall y \in Y & q(y) - \sum_\varepsilon \nu(y,\varepsilon) & = 0 \\
\forall \varepsilon \in \mathcal{E}^r & \sum_y \nu(y,\varepsilon) - f^r(\varepsilon;\theta_\varepsilon) & = 0 \\
 & \sum_{y,\varepsilon} \nu(y,\varepsilon) - 1 & = 0 \\
\forall i, y_i, y_i', \varepsilon_i & \sum_{y_{-i}} \sum_{\varepsilon_{-i}} \nu(y, \varepsilon_i, \varepsilon_{-i}) \left(\pi_i(y_i', y_{-i}, \varepsilon_i; x, \theta) - \pi_i(y, \varepsilon_i; x, \theta) \right) & \leq 0.
\end{array}$$

Although in $(P0)$ the minimum is taken over $q \in Q^{BCE}_\theta(x)$ only, here we minimize over both a vector of predictions $q \in \mathcal{P}_Y$ and a distribution $\nu \in \mathcal{P}_{Y \times \mathcal{E}^r}$ whose marginal on Y corresponds to q. The restriction that q must be a BCE prediction is now incorporated by imposing that ν must satisfy the constraints that characterize BCE distributions, as specified in Definition 2.

Step 2 - Vectorization: The discretized ν is a matrix with dimensions $|Y| \times r$; we define $v = \text{vec}(\nu)$, the vectorized ν that stacks the columns of ν in a vector with $d_\nu = |Y| \cdot r$ rows. We further specify how this vector is constructed; the vector v is a column vector formed by r columns

$$\nu_y(\varepsilon) = \begin{bmatrix} \nu(y^1,\varepsilon) \\ \nu(y^2,\varepsilon) \\ .. \\ .. \\ \nu(y^{|Y|},\varepsilon) \end{bmatrix}$$

[11] We have experimented with other discretization techniques (e.g. Halton sets, random draws) and have found negligible impact on our results as long as $\mathcal{E}^r$ includes at least some relatively extreme (both postive and negative) payoff types. This is because the incentive compatibility constraint of BCE is more likely to be binding for these values.

where vectors of actions are orderer in an order $y^1 ... y^{|Y|}$, and for a specific value of ε. The full v is then:

$$v = \begin{bmatrix} \nu_y\left(\varepsilon^1\right) \\ \nu_y\left(\varepsilon^2\right) \\ .. \\ .. \\ \nu_y\left(\varepsilon^r\right) \end{bmatrix}.$$

To define orderings for both ε and y, we start from complete orderings of both $\mathcal{E}_i$ and Y_i, summarized by the respective sets of indices. Then, we order the vectors as follows:

$$\begin{aligned}
\varepsilon^1 &= \left(\varepsilon_1^1, ..., \varepsilon_{|N|}^1\right); y^1 = \left(y_1^1, ..., y_{|N|}^1\right) \\
\varepsilon^2 &= \left(\varepsilon_1^1, ..., \varepsilon_{|N|-1}^1, \varepsilon_{|N|}^2\right); \varepsilon^2 = \left(y_1^1, ..., y_{|N|-1}^1, y_{|N|}^2\right) \\
&.... \\
\varepsilon^{r_{|N|}} &= \left(\varepsilon_1^1, ..., \varepsilon_{|N|-1}^1, \varepsilon_{|N|}^{r_{|N|}}\right); y^{r_{|N|}} = \left(y_1^1, ..., y_{|N|-1}^1, y_{|N|}^{r_{|N|}}\right) \\
\varepsilon^{r_{|N|}+1} &= \left(\varepsilon_1^1, ..., \varepsilon_{|N|-1}^2, \varepsilon_{|N|}^1\right); y^{r_{|N|}+1} = \left(y_1^1, ..., y_{|N|-1}^2, y_{|N|}^1\right) \\
\varepsilon^{r_{|N|}+2} &= \left(\varepsilon_1^1, ..., \varepsilon_{|N|-1}^2, \varepsilon_{|N|}^2\right); y^{r_{|N|}+2} = \left(y_1^1, ..., y_{|N|-1}^2, y_{|N|}^2\right) \\
&..
\end{aligned}$$

We then transform $(P1)$ by defining new variables $\tilde{p} = \hat{P}_{n,x} - q$ and

$$z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} \tilde{p} \\ v \end{bmatrix}.$$

As the set of predictions is a subset of the $(|Y|-1)$-dimensional simplex, we modify the objective of the program to $(\tilde{b}, 0)^\top \left(\hat{P}_{n,x} - q\right)$, where $\tilde{b}$ is a vector in the $(|Y|-1)-$dimensional closed ball $\mathbb{B}_{n,x}$. This modified objective yields a value of zero if and only if the original program has a value of zero. The transformed program is:

$$\begin{aligned}
\max_{\tilde{b}\in\mathbb{R}^{|Y|-1}} \min_{z_1\in\mathbb{R}^{|Y|}, z_2\in\mathbb{R}_+^{d_\nu}} \quad & \begin{bmatrix} \tilde{b} \\ 0_{d_\nu+1} \end{bmatrix}^\top z, \quad && (P2) \\
s.t. \quad & \tilde{b}^\top W_{n,x} \tilde{b} && \leq 1 \\
& A_{eq} z && = a \\
& A_{ineq} z && \leq 0_{d_{ineq}},
\end{aligned}$$

where A_{eq}, A_{ineq} and are matrices that stack, respectively, linear equality constraints and linear inequalities, and whose numbers of rows are $d_{eq} = |Y| + r + 1$ and $d_{ineq} = \sum_{i\in N} (|Y_i| \cdot |Y_i - 1| \cdot r_i)$. The object a is a vector of constants, and we use 0_d, 1_d and I_d to denote the $d-$vector of zeros and ones, and the $d \times d$

identity matrix. To construct the matrix A_{eq}, notice that the equality constraints in $(P1)$ can be written as

$$\begin{aligned} I_{|Y|}z_1 + A_{eq}^1 z_2 &= \hat{P}_{n,x} \\ A_{eq}^2 z_2 &= f^r(\theta_\varepsilon) \\ 1_{d_v}^\top z_2 &= 1, \end{aligned}$$

where A_{eq}^1 is a matrix with $|Y|$ rows whose $k-$th row is made of r copies of a $|Y|-$vector with components all zeros except for the $k-$th component that is equal to one, and A_{eq}^2 is a block-diagonal matrix with r rows and $1_{|Y|}^\top$ on the diagonal. The $d_{eq} \times d_z$ matrix A_{eq} is then

$$A_{eq} = \begin{bmatrix} I_{|Y|} & A_{eq}^1 \\ 0_{(r\cdot|Y|)} & A_{eq}^2 \\ 0_{|Y|}^\top & 1_{d_v}^\top \end{bmatrix}$$

with $d_z = |Y| \cdot (r+1)$; a is a $d_{eq}-$vector defined as

$$a = \begin{bmatrix} \hat{P}_{n,x} \\ f^r(\theta_\varepsilon) \\ 1 \end{bmatrix}.$$

The incentive compatibility inequality constraints in $(P1)$ are also linear, so that the matrix A_{ineq} can be constructed in a similar way.

Step 3 - Duality and Maximization Program: Although $(P2)$ is in the form of a maxmin problem, it can be transformed into a maximization problem by considering the dual of the inner minimization:

$$\begin{aligned} \max_{\tilde{b}\in\mathbb{R}^{|Y|-1},\lambda_{eq}\in\mathbb{R}^{d_{eq}},\lambda_{ineq}\in\mathbb{R}_+^{d_{ineq}}} \quad & -\begin{bmatrix} a \\ 0_{d_{ineq}} \end{bmatrix}^\top \begin{bmatrix} \lambda_{eq} \\ \lambda_{ineq} \end{bmatrix} && (P3) \\ s.t. \quad & \tilde{b}^\top W_{n,x}\tilde{b} \leq 1 \\ & \left(A^\top\right)_{1:|Y|} \begin{bmatrix} \lambda_{eq} \\ \lambda_{ineq} \end{bmatrix} = -\begin{bmatrix} \tilde{b} \\ 0 \end{bmatrix} \\ & \left(A^\top\right)_{|Y|+1:d_z} \begin{bmatrix} \lambda_{eq} \\ \lambda_{ineq} \end{bmatrix} \geq 0_{d_\nu}, \end{aligned}$$

where $A = \begin{bmatrix} A_{eq} \\ A_{ineq} \end{bmatrix}$, the vectors λ_{eq} and λ_{ineq} are the dual variables associated to the constraints of $(P2)$, and $\left(A^\top\right)_{1:|Y|}$ and $\left(A^\top\right)_{|Y|+1:d_z}$ denote the first $|Y|$ and the last rows of the matrix $A^\top$ and $d_A = d_{eq} + d_{ineq}$ is the number of rows of A. By strong duality, as well as by the existence of BCE, $(P3)$ has the same value than $(P2)$.

Finally, let $w = (\tilde{b}^\top, \lambda_{eq}^\top, \lambda_{ineq}^\top)^\top$, $\gamma = (0_{|Y|-1}^\top, a^\top, 0_{d_{ineq}}^\top)^\top$, and

$$\Gamma_1 = \begin{bmatrix} W_{n,x} & 0_{|Y|-1\times d_A} \\ 0_{d_Z\times|Y|-1} & 0_{d_A\times d_A} \end{bmatrix},\ \Gamma_2 = \begin{bmatrix} I_{|Y|-1} & \left(A^\top\right)_{1:|Y|} \\ 1_{|Y|}^\top & \end{bmatrix},\ \Gamma_3 = \begin{bmatrix} 0_{d_\nu\times|Y|-1} & \left(A^\top\right)_{|Y|+1:d_z} \end{bmatrix}.$$

Then one may represent $(P3)$ as in (19).

A.2 Computing $\sup_\theta p_n(\theta)$

Below, we drop the subscript n from p_n as it does not play a role. The following algorithm yields a sequence of tentative optimal values $p(\theta_L^*), L = k+1, k+2, \ldots,$ which tends to $\sup_{\theta\in\Theta} p(\theta)$.

Step 1 : Draw randomly (uniformly) over Θ a set $(\theta^{(1)}, \ldots, \theta^{(k)}$ of initial evaluation points. Evaluate $p(\theta^{(\ell)}), \ell = 1, \ldots, k-1$. Initialize $L = k$.

Step 2 : Record the tentative optimal value

$$p(\theta_L^*) = \max\{p(\theta^{(\ell)}), \ell \in \{1, \ldots, L\}\}. \tag{25}$$

If $p(\theta_L^*) > \alpha$, halt the algorithm. Set $\phi = 0$ (i.e. do not reject H_0). Otherwise, proceed to Step 3.

Step 3 : Approximate $\theta \mapsto p(\theta)$ by a flexible auxiliary model. A commonly used choice is a Gaussian-process regression, which models each observed value $Y^{(\ell)} = p(\theta^{(\ell)})$ as

$$Y^{(\ell)} = \mu + \zeta(\theta^{(\ell)}),\ \ell = 1, \ldots, L, \tag{26}$$

where ζ is a mean-zero Gaussian process indexed by θ with constant variance and a correlation

$$\mathrm{Corr}(\zeta(\theta), \zeta(\theta')) = K_\beta(\theta - \theta') \tag{27}$$

for some kernel function K_β (e.g. $K_\beta = \exp(-\sum_{h=1}^d (\theta_h - \theta_h')^2/\beta_h)$). The unknown parameters can be estimated by running a feasible-GLS regression of $\mathbf{Y} = (Y^{(1)}, \ldots, Y^{(L)})$ on a constant with the given correlation matrix. The (best linear) predictor of $p(\theta)$ then has a closed form[12]

$$p_L(\theta) = \hat{\mu} + \mathbf{r}_L(\theta)^\top \mathbf{R}_L^{-1}(\mathbf{Y} - \hat{\mu}\mathbf{1}). \tag{28}$$

This predictor coincides with p at the evaluation points (i.e., $p_L(\theta^{(\ell)}) = p(\theta^{((\ell))}), \ell = 1, \ldots, L$) providing an analytical interpolation. Also, the uncertainty left in p is quantified by the following variance:

$$\hat{\varsigma}^2 s_L^2(\theta) = \hat{\varsigma}^2 \left(1 - \mathbf{r}_L(\theta)^\top \mathbf{R}_L^{-1}\mathbf{r}_L(\theta) + \frac{(1 - \mathbf{1}^\top \mathbf{R}_L^{-1}\mathbf{r}_L(\theta))^2}{\mathbf{1}^\top \mathbf{R}_L^{-1}\mathbf{1}}\right). \tag{29}$$

Step 4: With probability $1-\epsilon$, obtain the next evaluation point $\theta^{(L+1)}$ as

$$\theta^{(L+1)} \in \arg\max_{\theta\in\Theta} (p_L(\theta) - p(\theta_L^*))\Phi\Big(\frac{p_L(\theta) - p(\theta_L^*)}{\hat{\varsigma} s_L(\theta)}\Big) + \hat{\varsigma} s_L(\theta)\phi\Big(\frac{p_L(\theta) - p(\theta_L^*)}{\hat{\varsigma} s_L(\theta)}\Big), \tag{30}$$

where the objective function is called the *expected improvement function.*

Set $L \leftarrow L+1$ and return to Step 1. Repeat the steps until convergence. The algorithm can be interpreted as follows. This algorithm first evaluates p on a coarse grid (Step 1). It then approximates p by a tractable

[12] Its derivative also has a closed form. See Jones et al. (1998).

Gaussian process regression model, which can be used to guide the determination of the next evaluation point to draw. For this, we need to take into account where the maximum is likely to lie (exploitation) and where the current approximation is rough and benefit from additional evaluation points (exploration). The expected improvement function in Step 4 provides a criterion for how to trade them off optimally. Repeating this process generates a sequence of that tends to the global maximum of p. This can be implemented by open softwares such as DACE (in Matlab) and MOE (in Python).[13]

B Proofs

B.1 Definitions and Notation

Below, we introduce objects that will be used in the proof of auxiliary lemmas. In Lemma 4 below, we consider a sequence η_n such that $\tau_n^{-1}\eta_n \to \pi \in \mathbb{R}^{K\times X}_{-,\infty}$. For $\xi = \tau_n^{-1}\eta_n(b,x)$, define

$$\varphi(\xi) \equiv \begin{cases} 0 & \text{if } \xi \geq -1 \\ -\infty & \text{if } \xi < -1 \ , \end{cases} \qquad \varphi^*(\xi) \equiv \begin{cases} \varphi(\xi) & \text{if } \pi = 0 \\ -\infty & \text{if } \pi < 0 \ . \end{cases} \tag{31}$$

Recall that W_x is the probability limit of $W_{n,x}$ (see Assumption 4). Let $K \subset \mathbb{R}^{|Y|}$ be a compact set such that $\mathbb{B}_x \subset K^{-\epsilon}$ uniformly across DGPs for some uniform constant $\epsilon > 0$, where $K^{-\epsilon} = \{x \in K : \inf_{y\in K^c} \|x - y\| \geq \epsilon\}$. We then define an empirical process $\mathbb{G}_n$ (indexed by $(b,x) \in K \times X$) and bootstrapped empirical process $\mathbb{G}_n^*$ as follows:

$$\mathbb{G}_n(b,x) \equiv \sqrt{n}b^\top(\hat{P}_{n,x} - P_{y|x}) \ , \qquad \mathbb{G}_n^*(b,x) \equiv \sqrt{n}b^\top(\hat{P}^*_{n,x} - \hat{P}_{n,x}). \tag{32}$$

Under our assumptions $\mathbb{G}_n$ converges weakly (in the sense of Hoffmann-Jørgensen) to a tight Gaussian process $\mathbb{G}$ (van der Vaart and Wellner, 1996). For any $\pi(b,x) \in [-\infty, 0]$, let

$$\pi^*(b,x) = \begin{cases} 0 & \text{if } \pi(b,x) = 0 \\ -\infty & \text{if } \pi(b,x) < 0. \end{cases} \tag{33}$$

Define $c_{\pi^*}(1-\alpha)$ be the $1-\alpha$ quantile of

$$\sup_{(b,x)\in\Psi_\infty} \mathbb{G}(b,x), \tag{34}$$

where

$$\Psi_\infty = \{(b,x) \in K \times X : \pi^*(b,x) = 0,\ b \in \mathbb{B}_x\}. \tag{35}$$

This is the set of constraints that will be selected by the GMS asymptotically.

Let $\mathcal{F} \equiv \{(\theta, P) : P \in \mathcal{P}_\theta,\ \theta \in \Theta\}$ be the set of (θ, P) pairs that are compatible with our assumptions. Following Andrews and Soares (2010), we introduce a one-to-one mapping between $(\theta, P) \in \mathcal{F}$ and a new

[13] These softwares are available from http://www.omicron.dk/dace.html (DACE) and https://github.com/wujian16/Cornell-MOE (MOE) respectively.

parameter $\gamma = (\gamma_1, \gamma_2, \gamma_3)$ with corresponding parameter space Γ. For each $(b, x) \in K \times X$, $\gamma_1 \in \mathbb{R}_-^{K \times X}$ is defined by the relation

$$b^\top P_{y|x} - h(b, Q_{\theta,S^r}^{BCE}(x)) - \gamma_1(b, x) = 0. \tag{36}$$

For each $x \in X$, let $W_x = \text{AsyVar}_P(\sqrt{n}\hat{P}_{n,x})$, and we let $\gamma_2 = (\theta, vech(W_{x_1}), \ldots, vech(W_{x_{|X|}}))^\top$. Finally, we let $\gamma_3 = P$. We then define Γ as

$$\Gamma \equiv \big\{\gamma = (\gamma_1, \gamma_2, \gamma_3) : \text{ for some } (\theta, P) \in \mathcal{F}, \gamma_1 \text{ satisfies } (36), \\ \gamma_2 = (\theta, vech(W_{x_1}), \ldots, vech(W_{x_{|X|}}))^\top,\ \gamma_3 = P\big\}. \tag{37}$$

Note that, since we impose $(\theta, P) \in \mathcal{F}$, all parameters in Γ must respect Assumptions 2-5. In what follows, for any sequence $\{\gamma_n\}$, we let

$$\eta_n(b, x) \equiv \sqrt{n}\gamma_{1,n}(b, x). \tag{38}$$

Other notation and definitions used throughout are collected in the following table.

$a \lesssim b$	$a \leq Mb$ for some constant M.
$\|\cdot\|_{op}$	the operator norm for linear mappings.
$\|\cdot\|_{\mathcal{F}}$	the supremum norm over $\mathcal{F}$.
$\|\cdot\|_{\psi_2}$	the sub-Gaussian norm: $\|X\|_{\psi_2} = \inf\{t > 0 : E[\exp(X^2/t^2)] \leq 2\}$.
$N(\epsilon, \mathcal{F}, \|\cdot\|)$	covering number of size ϵ for $\mathcal{F}$ under norm $\|\cdot\|$.
$N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$	bracketing number of size ϵ for $\mathcal{F}$ under norm $\|\cdot\|$.
$X_n \overset{P^n}{\rightsquigarrow} X$	X_n weakly converges to X under $\{P^n\}$.

B.2 On Studentization

Lemma 3. *Let $\mathbb{B}_{n,x} = \{b \in \mathbb{R}^{|Y|} : \|b\|_{W_{n,x}} \leq 1\}$. Then (10) holds.*

Proof. Let $\mathbb{S}_n = \{b : \|b\|_{W_{n,x}} = 1\}$ be the set of unit vectors with respect to $\|\cdot\|_{W_{n,x}}$. Note that the map $b \mapsto b^\top \hat{P}_{n,x} - h(b, Q_{\theta,S^r}^{BCE}(x))$ is positively homogeneous because the support function is positively homogeneous (Molchanov, 2005, Appendix F). One may then write

$$\begin{aligned}\sup_{b \in \mathbb{B}_{n,x}} \sqrt{n}\{b^\top \hat{P}_{n,x} - h(b, Q_{\theta,S^r}^{BCE}(x))\} &= \sup_{\gamma \in [0,1]} \gamma \sup_{\bar{b} \in \mathbb{S}_n} \sqrt{n}\{\bar{b}^\top \hat{P}_{n,x} - h(\bar{b}, Q_{\theta,S^r}^{BCE}(x))\} \\ &= \begin{cases} \sup_{\bar{b} \in \mathbb{S}_n} \sqrt{n}\{\bar{b}^\top \hat{P}_{n,x} - h(\bar{b}, Q_{\theta,S^r}^{BCE}(x))\} & \text{if } \sup_{\bar{b} \in \mathbb{S}_n} \sqrt{n}\{\bar{b}^\top \hat{P}_{n,x} - h(\bar{b}, Q_{\theta,S^r}^{BCE}(x))\} > 0 \\ 0 & \text{if } \sup_{\bar{b} \in \mathbb{S}_n} \sqrt{n}\{\bar{b}^\top \hat{P}_{n,x} - h(\bar{b}, Q_{\theta,S^r}^{BCE}(x))\} \leq 0, \end{cases}\end{aligned} \tag{39}$$

where the second equality holds because it is optimal to set $\gamma = 0$ whenever the optimal value of the inner maximization problem is non-positive, and it is optimal to set $\gamma = 1$ otherwise. Finally, note that one may represent $\bar{b}$ as $\bar{b} = b/(b^\top W_{n,x} b)^{1/2}$ for a nonzero vector b. Therefore, again using the positive homogeneity, we may write

$$\sup_{b\in\mathbb{B}_{n,x}} \sqrt{n}\{b^\top \hat{P}_{n,x} - h(b, Q_{\theta,S^r}^{BCE}(x))\} = \sup_{b\in\mathbb{R}^{|Y|}\setminus\{0\}} \sqrt{n}\Big\{\frac{b^\top \hat{P}_{n,x} - h(b, Q_{\theta,S^r}^{BCE}(x))}{\sqrt{nb^\top W_{n,x} b}}\Big\}_+. \tag{40}$$

□

B.3 Results and Auxiliary Lemmas on Hypothesis Testing

Proof of Theorem 2. We first observe that rejection based on the p-value can be restated as follows:

$$p_n(\theta) < \alpha,\ \forall\theta\in\Theta, \quad \Leftrightarrow \quad T_n(\theta) > \hat{c}_n(\theta, 1-\alpha),\ \forall\theta\in\Theta, \tag{41}$$

where $\hat{c}_n(\theta, 1-\alpha) = \inf\{c : P^*(T_n^*(\theta) \leq c|y^n, x^n) \geq 1-\alpha\}$. Therefore, for any fixed $\theta \in \Theta$,

$$P\big(\sup_{\tilde{\theta}\in\Theta} p_n(\tilde{\theta}) < \alpha\big) \leq P(T_n(\theta) > \hat{c}_n(\theta, 1-\alpha)). \tag{42}$$

Hence,

$$\sup_{P\in\mathcal{P}} E_P[\phi] \leq \sup_{(\theta,P)\in\mathcal{F}} P(T_n(\theta) > \hat{c}_n(\theta, 1-\alpha)). \tag{43}$$

Let $\{(P_n, \theta_n) \in \mathcal{P}\times\Theta\}_{n=1}^\infty$ be a sequence such that

$$\limsup_{n\to\infty} P_n(T_n(\theta_n) > \hat{c}_n(\theta_n, 1-\alpha)) = \limsup_{n\to\infty} \sup_{(\theta,P)\in\mathcal{F}} P(T_n(\theta) > \hat{c}_n(\theta, 1-\alpha)). \tag{44}$$

Let $\mathrm{RP}_n(\gamma_n) = P_n(T_n(\theta_n) > \hat{c}_n(\theta_n, 1-\alpha))$, where γ_n is a sequence of parameters defined as in Section B.1. Let $\{u_n\}$ be a subsequence of $\{n\}$ such that $\lim_{n\to\infty} \mathrm{RP}_{u_n}(\gamma_{u_n})$ exists and

$$\lim_{n\to\infty} \mathrm{RP}_{u_n}(\gamma_{u_n}) = \limsup_{n\to\infty} \mathrm{RP}_n(\gamma_n). \tag{45}$$

Such a sequence exsits without loss of generality. Following (S1.1)-(S1.4) in Andrews and Soares (2010) (supplementary material), it is straightforward to constrcut a further subsequence $\{w_n\}$ of $\{u_n\}$ such that (i) $\eta_{w_n} = \sqrt{w_n}\gamma_{1,w_n} \to \eta \in \mathbb{R}^{K\times X}_{-,\infty}$, (ii) $\tau_{w_n}^{-1}\eta_{w_n} \to \pi \in \mathbb{R}^{K\times X}_{-,\infty}$, and $\gamma_{2,w_n} \to \gamma_2 \in \mathrm{cl}(\Gamma_2)$. By Lemma 4, it then follows that $\lim_{n\to\infty} \mathrm{RP}_{w_n}(\gamma_{w_n}) \leq \alpha$. Hence, the conclusion of the theorem holds. □

Lemma 4. *Suppose Assumptions 1-5 hold. Let $\{\gamma_n : n \geq 1\}$ be a sequence in Γ where the components η_n and $(W_{n,x}, x \in X)$ of $\gamma_{n,\eta}$ satisfy (i) $\eta_n \to \eta \in \mathbb{R}^{K\times X}_{-,\infty}$ (pointwise in (b, x)); (ii) $\tau_n^{-1}\eta_n \to \pi \in \mathbb{R}^{K\times X}_{-,\infty}$; (iii) $\gamma_{2,n} \to \gamma \in \mathbb{R}^{d_{\gamma_2}}_{\pm\infty}$, where d_{γ_2} is the dimension of γ_2. Then, (a)-(c) hold.*

(a) $\hat{c}_n(\theta_n, 1-\alpha) \geq c_n^*$, a.s. for all n for a sequence of random variables $\{c_n^*\}$ such that $c_n^* \xrightarrow{p} c_{\pi^*}(1-\alpha)$.

(b) $\limsup_{n\to\infty} P_{\eta_n}(\tilde{T}_{1,n}(\theta_n) > \hat{c}_n(\theta_n, 1-\alpha)) \leq \alpha$.

(c) Parts (a) and (b) hold with all subsequences $\{w_n\}$ in place of $\{n\}$ provided that (i)–(iii) hold with w_n in place of n.

Proof. Rewrite T_n and T_n^* as

$$T_n(\theta) = \sup_{x\in X}\sup_{b\in\mathbb{B}_{n,x}}\{\mathbb{G}_n(b,x)+\eta_n(b,x)\},$$
$$T_n^*(\theta) = \sup_{x\in X}\sup_{b\in\mathbb{B}_{n,x}}\{\mathbb{G}_n^*(b,x)+\varphi(\hat{\xi}_n(b,x;\theta))\},$$

where $\hat{\xi}_n(b,x;\theta)\equiv\tau_n^{-1}\hat{\eta}_n(b,x;\theta)$. Note that $c_{\pi^*}(1-\alpha)$ is the $1-\alpha$ quantile of

$$\sup_{x\in X}\sup_{b\in\mathbb{B}_x}\{\mathbb{G}(b,x)+\pi^*(b,x)\}.$$

To prove part (a), first, when $c_{\pi^*}(1-\alpha)=0$, define $c_n^*=0$ so that

$$\hat{c}_n(\theta_n,1-\alpha)\geq c_n^*\xrightarrow{p}c_{\pi^*}(1-\alpha)$$

trivially holds since $T_n^*(\theta)\geq 0$ for all $\theta\in\Theta$.

Next, suppose $c_{\pi^*}(1-\alpha)>0$. By construction,

$$\varphi^*(\hat{\xi}_n(b,x;\theta_n))\leq\varphi(\hat{\xi}_n(b,x;\theta_n)) \text{ a.s. } \forall n \text{ and } (b,x). \tag{46}$$

Let c_n^* be the $1-\alpha$ quantile of

$$\sup_{x\in X}\sup_{b\in\mathbb{B}_{n,x}}\{\mathbb{G}_n^*(b,x)+\varphi^*(\hat{\xi}_n(b,x;\theta))\}.$$

Then $\hat{c}_n(\theta_n,1-\alpha)\geq c_n^*$ a.s. $\forall n$ by (46).

We now show $c_n^*\xrightarrow{p}c_{\pi^*}(1-\alpha)>0$. Let E_{M_n} denote the expectation operator with respect to the distribution of the (multinomial) weights for bootstrap resampling and let BL_1 denote the set of functions on $K\times X$ that are boundedly Lipschitz with a Lipschitz constant of 1. By Lemma 8 and arguing as in the proof of Lemma 10, $\mathbb{G}_n^*$ is a bootstrapped empirical process defined on a uniform Donsker class. Therefore, applying Theorem 3.6.2 in van der Vaart and Wellner (1996) under any $\{P_n\}\subset\mathcal{P}$, we have

$$\sup_{h\in BL_1}|E_{M_n}[h(\mathbb{G}_n^*)|(y^n,x^n)]-E[h(\mathbb{G})]|\xrightarrow{P_n}0. \tag{47}$$

For each (b,x),

$$\begin{aligned}\hat{\xi}_n(b,x;\theta_n) &= \tau_n^{-1}\sqrt{n}(b^\top\hat{P}_{n,x}-h(b,Q^{BCE}_{\theta_n,S^r}(x)))\\ &= \tau_n^{-1}\sqrt{n}(b^\top P_{n,y|x}-h(b,Q^{BCE}_{\theta_n,S^r}(x)))+\tau_n^{-1}\sqrt{n}b^\top(\hat{P}_{n,x}-P_{n,y|x})\\ &= \tau_n^{-1}\eta_n(b,x)+o_{p^*}(1)\\ &\xrightarrow{p^*}\pi(b,x),\end{aligned} \tag{48}$$

where the convergence is by (ii) and the second equality is by Assumption 3 that guarantees $\hat{P}_{nx}-P_{n,y|x}=O_p(n^{-1/2})$ for each x, which in turn implies $\tau_n^{-1}\sqrt{n}(b^\top\hat{P}_{n,x}-P_{n,y|x})=o_{p^*}(1)$, and for r.v.'s z_n and z, the convergence $z_n\xrightarrow{p^*}z$ is defined as $P^*(|z_n-z|>\varepsilon|y^n,x^n)\xrightarrow{p}0$

We now show that $\varphi^*(\xi(b,x)) \to \varphi^*(\pi(b,x))$ for any sequence $\xi(b,x)$ for which $\xi(b,x) \to \pi(b,x)$. If $\pi(b,x) = 0$, then

$$\varphi^*(\xi(b,x)) = \varphi(\xi(b,x)) \to \varphi(\pi(b,x)) = \varphi^*(\pi(b,x)) = 0$$

as $\xi(b,x) \to \pi(b,x)$, by the definitions in (31). If $\pi(b,x) < 0$, then

$$\varphi^*(\xi(b,x)) = -\infty = \varphi^*(\pi(b,x))$$

by (31). Therefore, as $\xi(b,x) \to \pi(b,x)$, we have $\varphi^*(\xi(b,x)) \to \varphi^*(\pi(b,x)) = \pi^*(b,x)$ where the last equality is by the definitions in (31) and (33).

Let $\delta > 0$. Define the following sets of sample paths:

$$A_{1n} \equiv \Big\{(y^n, x^n) : \inf_{(b,x):\pi(b,x)=0} \hat{\xi}_n(b,x) < -1\Big\} \tag{49}$$

$$A_{2n} \equiv \Big\{(y^n, x^n) : \sup_{x\in X} d_H(\mathbb{B}_{n,x}, b_x) > \delta\Big\}. \tag{50}$$

Lemmas 5 and 7 show that $P_n(A_{jn}) \to 0$, for $j = 1, 2$.

Now note that, for any $a \in \mathbb{R}$,

$$\sup_{x\in X}\sup_{b\in\mathbb{B}_{n,x}} \{\mathbb{G}_n^*(b,x) + \varphi^*(\hat{\xi}_n(b,x;\theta_n))\} \le a$$
$$\Leftrightarrow \sup_{(b,x)\in\Psi_n} \{\mathbb{G}_n^*(b,x) + \varphi^*(\hat{\xi}_n(b,x;\theta_n))\} \le a \text{ and } \sup_{(b,x)\in\Psi_n^c} \{\mathbb{G}_n^*(b,x) + \varphi^*(\hat{\xi}_n(b,x;\theta_n))\} \le a, \tag{51}$$

where $\Psi_n = \{(b,x) \in K \times X : \pi^*(b,x) = 0,\ b \in \mathbb{B}_{n,x}\}$ and $\Psi_n^c = \{(b,x) \in K \times X : \pi^*(b,x) \neq 0,\ b \in \mathbb{B}_{n,x}\}$.

For $(b,x) \in \Psi_n^c$, we have previously shown that $\varphi^*(\hat{\xi}_n(b,x;\theta_n)) = -\infty$, and thus $\sup_{(b,x)\in\Psi_n^c}\{\mathbb{G}_n^*(b,x) + \varphi^*(\hat{\xi}_n(b,x;\theta_n))\} = 0$. Therefore, by (51), one may write

$$\begin{aligned} &P_{M_n}\Big(\sup_{x\in X}\sup_{b\in\mathbb{B}_{n,x}} \{\mathbb{G}_n^*(b,x) + \varphi^*(\hat{\xi}_n(b,x;\theta_n))\} \le a \big| y^n, x^n\Big) \\ &= P_{M_n}\Big(\sup_{(b,x)\in\Psi_n} \{\mathbb{G}_n^*(b,x) + \varphi^*(\hat{\xi}_n(b,x;\theta_n))\} \le a, 0 \le a \big| y^n, x^n\Big) \\ &= \begin{cases} P_{M_n}\Big(\sup_{(b,x)\in\Psi_n} \{\mathbb{G}_n^*(b,x) + \varphi^*(\hat{\xi}_n(b,x;\theta_n))\} \le a \big| y^n, x^n\Big) & \text{if } a \ge 0, \\ 0 & \text{if } a < 0, \end{cases}. \end{aligned} \tag{52}$$

Similarly, for any $a \in \mathbb{R}$,

$$\sup_{x\in X}\sup_{b\in\mathbb{B}_x} \{\mathbb{G}(b,x) + \pi^*(b,x)\} \le a$$
$$\Leftrightarrow \sup_{(b,x)\in\Psi_\infty} \{\mathbb{G}(b,x) + \pi^*(b,x)\} \le a \text{ and } \sup_{(b,x)\in\Psi_\infty^c} \{\mathbb{G}(b,x) + \pi^*(b,x)\} \le a. \tag{53}$$

Observe that $\pi^*(b,x) - \infty$ for any $(b,x) \in \Psi_\infty^c$ implying $\sup_{(b,x)\in\Psi_\infty^c}\{\mathbb{G}(b,x) + \pi^*(b,x)\} = 0$. Therefore, by mimicking the argument in (52),

$$P\Big(\sup_{x\in X}\sup_{b\in\mathbb{B}_x}\{\mathbb{G}(b,x)+\pi^*(b,x)\}\le a\Big)=\begin{cases}P\Big(\sup_{(b,x)\in\Psi_\infty}\{\mathbb{G}(b,x)+\pi^*(b,x)\}\le a\Big) & \text{if } a\ge 0,\\ 0 & \text{if } a<0\end{cases}$$
$$=\begin{cases}P\Big(\sup_{(b,x)\in\Psi_\infty}\{\mathbb{G}(b,x)\}\le a\Big) & \text{if } a\ge 0,\\ 0 & \text{if } a<0,\end{cases}\tag{54}$$

where the last equality follows from $\pi^*(b,x)=0$ on Ψ_∞.

Now, for each $a\in\mathbb{R}$, let

$$r_n(a)\equiv P_{M_n}\Big(\sup_{x\in X}\sup_{b\in\mathbb{B}_{n,x}}\{\mathbb{G}_n^*(b,x)+\varphi^*(\hat{\xi}_n(b,x;\theta_n))\}\le a|(y^n,x^n)\Big)$$
$$-P\Big(\sup_{x\in X}\sup_{b\in\mathbb{B}_x}\{\mathbb{G}(b,x)+\pi^*(b,x)\}\le a\Big).\tag{55}$$

Suppose $(y^n,x^n)\in A_{1n}^c\cap A_{2n}^c$. On this event, $\varphi^*(\hat{\xi}_n(b,x;\theta_n))=0$ uniformly on Ψ_n, and thus $\sup_{\Psi_n}\{\mathbb{G}_n^*(b,x)+\varphi^*(\hat{\xi}_n(b,x;\theta_n))\}=\sup_{\Psi_n}\{\mathbb{G}_n^*(b,x)\}$. This, together with (52) and (54) imply that, for any given $a>0$ and any $\epsilon>0$,

$$P_n(\{|r_n(a)|>\epsilon\}\cap A_{1n}^c\cap A_{2n}^c)$$
$$=P_n\Big(\Big\{\big|P_{M_n}\big(\sup_{(b,x)\in\Psi_n}\{\mathbb{G}_n^*(b,x)\}\le a\big)-P\big(\sup_{(b,x)\in\Psi_\infty}\{\mathbb{G}(b,x)\}\le a\big)\big|>\epsilon\Big\}\cap A_{1n}^c\cap A_{2n}^c\Big)\to 0,\tag{56}$$

where the convergence to 0 follows from Lemma 11. Finally, By Lemma 5, we then have

$$P_n(|r_n(a)|>\epsilon)\le P_n(\{|r_n(a)|>\epsilon\}\cap A_{1n}^c\cap A_{2n}^c)+P_n(A_{1n}\cup A_{2n})\to 0.\tag{57}$$

Since this holds for all a in the neighborhood of $c_{\pi^*}(1-\alpha)>0$, by Lemma 5 of Andrews and Guggenberger (2010b), we have

$$c_n^*\xrightarrow{p}c_{\pi^*}(1-\alpha).$$

We now prove part (b). First, if $\pi(b,x)<0$, then $\eta(b,x)=\pi^*(b,x)=-\infty$ by conditions (i) and (ii), and if $\pi(b,x)=0$, then $\eta(b,x)\le 0$ and $\pi^*(b,x)=0$. That is, for each (b,x), we have $\pi^*(b,x)\ge\eta(b,x)$, and therefore

$$\sup_{x\in X}\sup_{b\in\mathbb{B}_x}\{\mathbb{G}(b,x)+\pi^*(b,x)\}_+\ge\sup_{x\in X}\sup_{b\in\mathbb{B}_x}\{\mathbb{G}(b,x)+\eta(b,x)\}_+$$

and thus

$$c_{\pi^*}(1-\alpha)\ge c_\eta(1-\alpha),\tag{58}$$

where $c_\eta(1-\alpha)$ denotes the $1-\alpha$ quantile of $\sup_{x\in X}\sup_{b\in\mathbb{B}_x}\{\mathbb{G}(b,x)+\eta(b,x)\}$.

Next, under $\{\gamma_{n,\eta}:n\ge 1\}$, $\mathbb{G}_n\rightsquigarrow\mathbb{G}$ and $\eta_n\rightsquigarrow\eta$ by a parallel argument to that for the bootstrap weak convergence above. Then, by the continuous mapping theorem, we have

$$T_n(\theta_n) \xrightarrow{d} J_\eta, \tag{59}$$

where J_η is the distribution of $\sup_{(b,x)\in B\times X}\{\mathbb{G}(b,x)+\eta(b,x)\}_+$. We then have

$$\limsup_{n\to\infty} P_{\gamma_{n,\eta}}(T_n(\theta_n) > \hat{c}_n(\theta_n, 1-\alpha)) \leq \limsup_{n\to\infty} P_{\gamma_{n,\eta}}(T_n(\theta_n) > c_n^*) \leq 1 - J_\eta(c_{\pi^*}(1-\alpha)-),$$

where the first inequality is by part (a) that $\hat{c}_n(\theta_n, 1-\alpha) \geq c_n^*$ and the second inequality is by part (a) that $c_n^* \xrightarrow{p} c_{\pi^*}(1-\alpha)$ and (59) with $J_\eta(a-)$ being the limit from the left of $J_\eta(\cdot)$ at a.

Suppose $c_{\pi^*}(1-\alpha) > 0$. Then

$$J_\eta(c_{\pi^*}(1-\alpha)-) = J_\eta(c_{\pi^*}(1-\alpha)) \geq 1-\alpha,$$

where the equality holds by $J_\eta(a)$ being continuous $\forall a > 0$ and the inequality is by (58). This proves part (b) for this case.

Now suppose $c_{\pi^*}(1-\alpha) = 0$. This implies that $c_\eta(1-\alpha) = 0$ by (58). Under $\{\gamma_{n,\eta} : n \geq 1\}$, we have

$$\limsup_{n\to\infty} P_{\gamma_{n,\eta}}(T_n(\theta_n) > 0) = 1 - J_\eta(0) = 1 - J_\eta(c_\eta(1-\alpha)) \leq \alpha$$

where the first equality is by (59). This proves part (b) for this case.

The proof of part (c) is analogous to that for parts (a) and (b) with w_n in place of n.

□

Let $I_x = (1\{y_\ell = y_1, x_\ell = x\}, \ldots, 1\{y_\ell = y_{|Y|}, x_\ell = x\}) \in \{0,1\}^{|Y|}$. Suppose for the moment p_x is known. We may then write $\mathbb{G}_n(b,x) = -\frac{1}{\sqrt{n}}\sum_{\ell=1}^n (Z_\ell(b) - E[Z_\ell(b)])/p_x$, where $Z_\ell(b) = -I_x' b$.

Lemma 5. *Suppose Assumptions 1-3 and 5 hold. Let $\tau_n \to \infty$. Then, for any sequence $\{P_n\} \subset \mathcal{P}$, $P_n(A_n) \to 0$.*

Proof. For each $x \in X$, let $\Pi(x) \equiv \{b \in \mathbb{B}_x : \pi(b,x) = 0\}$. Fix $b \in \Pi(x)$. Observe that

$$\begin{aligned}
\hat{\xi}_n(b,x) < -1 \ &\Leftrightarrow \tau_n^{-1}\hat{\eta}_n(b,x) < -1 \\
&\Leftrightarrow \tau_n^{-1}\eta_n(b,x) + \tau_n^{-1}\mathbb{G}_n(b,x) < -1 \\
&\Leftrightarrow \mathbb{G}_n(b,x) < -\tau_n - \eta_n(b,x) \\
&\Leftrightarrow \frac{1}{\sqrt{n}}\sum_{\ell=1}^n Z_\ell(b) - E[Z_\ell(b)] > p_x(\tau_n + \eta_n(b,x)) \\
&\Rightarrow \frac{1}{\sqrt{n}}\sum_{\ell=1}^n Z_\ell(b) - E[Z_\ell(b)] > \underline{\zeta}(\tau_n + \eta_n(b,x)),
\end{aligned}$$

where the last implication is due to $p_x \geq \underline{\zeta}$ for all x.

Let $\tilde{\tau}_n \equiv \tau_n + \inf_{b\in\Pi(x)} \eta_n(b,x)$ and note that $\tilde{\tau}_n \to \infty$.[14] Then, for any $x \in X$,

$$P\Big(\inf_{b\in\Pi(x)} \hat{\xi}_n(b,x) < -1\Big) \le P\Big(\sup_{b\in\Pi(x)} \frac{1}{\sqrt{n}}\sum_{\ell=1}^{n} Z_\ell(b) - E[Z_\ell(b)] > \underline{\zeta}\tilde{\tau}_n\Big) \tag{60}$$

$$= P\Big(\sup_{b\in\Pi(x)} \mathcal{Z}_n(b) > \underline{\zeta}\tilde{\tau}_n\Big) \tag{61}$$

$$= P\Big(\sup_{b\in\Pi(x)} \{\mathcal{Z}_n(b) - \mathcal{Z}_n(0)\} > \underline{\zeta}\tilde{\tau}_n\Big), \tag{62}$$

where $\mathcal{Z}_n(0) = 0$ by definition. By Lemma 6, $\mathcal{Z}_n$ is a sub-Gaussian process. The chaining tail inequality (van Handel, 2018, Theorem 5.29) then implies, for any $u \ge 0$,

$$P\Big(\sup_{b\in\Pi(x)} \{\mathcal{Z}_n(b) - \mathcal{Z}_n(0)\} > C\int_0^\infty \sqrt{\ln N(B,\|\cdot\|,\epsilon)}d\epsilon + u\Big) \le C\exp(-\frac{u^2}{CD^2}), \tag{63}$$

where C is a universal constant, $D = \text{diam}(B)$, and $N(B,\|\cdot\|,\epsilon)$ is the covering number of B. Note that

$$M \equiv \int_0^\infty \sqrt{\ln N(B,\|\cdot\|,\epsilon)}d\epsilon \le \int_0^D \sqrt{|Y|\ln(D/\epsilon)}d\epsilon < \infty, \tag{64}$$

where we used $N(B,\|\cdot\|,\epsilon) \le (\text{diam}(B)/\epsilon)^{|Y|}$, and $\ln N(B,\|\cdot\|,\epsilon) = 0$ for all $\epsilon > \text{diam}(D)$.

Let $u_n = \underline{\zeta}\tilde{\tau}_n - M$ and observe that $u_n > 0$ for all $n \ge \bar{N}$ sufficiently large. By (62)-(64), for any $n \ge \bar{N}$,

$$P\Big(\sup_{b\in\Pi(x)} \{\mathcal{Z}_n(b) - \mathcal{Z}_n(0)\} > \underline{\zeta}\tilde{\tau}_n\Big) \le C\exp(-\frac{u_n^2}{CD^2}) \tag{65}$$

By (60)-(62) and the union bound, we obtain

$$P(A_n) \le \sum_{x\in X} P\Big(\inf_{b\in\Pi(x)} \hat{\xi}_n(b,x) < -1\Big) \le \sum_{x\in X} C\exp(-\frac{u_n^2}{CD^2}). \tag{66}$$

The claim of the lemma then follows due to $u_n \to \infty$ and C, D being independent of P. $\square$

Lemma 6. *For each $b \in \mathbb{B}_x$, let $Z_\ell(b) = -I_x'b$ and let $\mathcal{Z}_n(b) = \frac{1}{\sqrt{n}}\sum_{\ell=1}^n Z_\ell(b) - E[Z_\ell(b)]$. Then, there exists $K \ge 0$ such that*

$$\|\mathcal{Z}_n(b) - \mathcal{Z}_n(\tilde{b})\|_{\psi_2} \le K\|b - \tilde{b}\|, \quad \textit{for all } b, \tilde{b} \in \mathbb{B}_x. \tag{67}$$

Proof. Note that

$$Z_\ell(b) - Z_\ell(\tilde{b}) \le |Z_\ell(b) - Z_\ell(\tilde{b})| \le \|I_x\|\|b - \tilde{b}\|, \tag{68}$$

and $W = \|I_x\|$ is a bounded random variable, which implies $\|W\|_{\psi_2} \le \frac{1}{\sqrt{\ln 2}}\|W\|_\infty$ (Vershynin, 2018, p.25). The inequality in (68) also implies $E[Z_\ell(b)] - E[Z_\ell(\tilde{b})] \le E[|W|]\|b-\tilde{b}\| \le \|W\|_\infty\|b-\tilde{b}\|$. Since $\|W\|_\infty \le \sqrt{|Y|}$,

[14] While $\tilde{\tau}_n$ depends on x, it does not create an issue. For notational simplicity, we will be implicit about $\tilde{\tau}_n$'s dependence on x below.

the triangle inequality implies

$$\big\|(Z_\ell(b) - E[Z_\ell(b)]) - (Z_\ell(\tilde{b}) - E[Z_\ell(\tilde{b})])\big\|_{\psi_2} \leq \sqrt{|Y|}(1 + \frac{1}{\sqrt{\ln 2}})\|b - \tilde{b}\|. \tag{69}$$

Note that $\mathcal{Z}_n(b) - \mathcal{Z}_n(\tilde{b})$ is the sum of independent mean-zero sub-Gaussian random variables $(Z_\ell(b) - E[Z_\ell(b)]) - (Z_\ell(\tilde{b}) - E[Z_\ell(\tilde{b})])$ and hence by Proposition 2.6.1 in Vershynin (2018),

$$\|\mathcal{Z}_n(b) - \mathcal{Z}_n(\tilde{b})\|_{\psi_2}^2 = \frac{1}{n}\Big\|\sum_{\ell=1}^{n}(Z_\ell(b) - E[Z_\ell(b)]) - (Z_\ell(\tilde{b}) - E[Z_\ell(\tilde{b})])\Big\|_{\psi_2}^2 \tag{70}$$

$$\leq \frac{1}{n}C\sum_{\ell=1}^{n}\big\|(Z_\ell(b) - E[Z_\ell(b)]) - (Z_\ell(\tilde{b}) - E[Z_\ell(\tilde{b})])\big\|_{\psi_2}^2 \tag{71}$$

$$\leq C\big(\sqrt{|Y|}(1 + \frac{1}{\sqrt{\ln 2}})\|b - \tilde{b}\|\big)^2, \tag{72}$$

where $C > 0$ is a universal constant. The claim then follows with $K = \sqrt{C}(1 + \frac{1}{\sqrt{\ln 2}})$. □

Lemma 7. *Suppose Assumptions 1-4 hold. Then, for any $\epsilon > 0$ and $\eta > 0$, there exists $\bar{N}_{\epsilon,\eta} \in \mathbb{N}$ such that*

$$\sup_{P\in\mathcal{P}} P\big(\sup_{x\in X} d_H(\mathbb{B}_{n,x}, \mathbb{B}_x) > \epsilon\big) < \eta. \tag{73}$$

for all $n \geq \bar{N}_{\epsilon,\eta}$.

Proof. We show convergence using the following isometry (Li et al., 2002, Theorem 1.1.12):

$$d_H(\mathbb{B}_{n,x}, \mathbb{B}_x) = \sup_{p\in\mathbb{S}^{|Y|-1}} |h(p, \mathbb{B}_{n,x}) - h(p, \mathbb{B}_x)|. \tag{74}$$

For this, recall that the support function of $\mathbb{B}_x$ is given by

$$h(p, \mathbb{B}_x) = \sup_{b\in\mathbb{B}_x} p^\top b = \sup p^\top b. \quad s.t.\ b^\top W_x b \leq 1 \tag{75}$$

Solving the quadratic program above yields

$$h(p, \mathbb{B}_x) = (p^\top W_x^{-1} p)^{1/2}, \tag{76}$$

where W_x^{-1} is well defined by $\underline{\lambda}(W_x) > 0$. A similar argument can show

$$h(p, \mathbb{B}_{n,x}) = (p^\top W_{n,x}^{-1} p)^{1/2}. \tag{77}$$

By the Cauchy-Schwarz inequality and p being a unit vector, it follows that

$$\begin{aligned}|h(p,\mathbb{B}_{n,x})^2-h(p,\mathbb{B}_x)^2| &= |p^\top(W_{n,x}^{-1}-W_x^{-1})p| \\ &= |p^\top W_{n,x}^{-1}(W_x-W_{n,x})W_x^{-1}p| \\ &\le \|p^\top W_{n,x}^{-1}\|\|(W_x-W_{n,x})W_x^{-1}p\| \\ &\le \|W_{n,x}^{-1}\|_{op}\|(W_x-W_{n,x})\|_{op}\|W_x^{-1}p\| \\ &\le \|W_{n,x}^{-1}\|_{op}\|W_x^{-1}\|_{op}\|W_x-W_{n,x}\|_{op}. \end{aligned} \tag{78}$$

Note that $\|W_x^{-1}\|_{op}=\underline{\lambda}(W_x)^{-1}\le\underline{\kappa}^{-1}$. Let $\varepsilon=\underline{\kappa}^{-1}-\underline{\lambda}(W_x)^{-1}$. By the Lipschitz continuity of $\underline{\lambda}$ for Hermitian matrices (Bhatia, 1997, Corollary III.2.6), there is $\delta>0$ such that

$$\|W_x-W_{n,x}\|_{op}\le\delta \;\Rightarrow\; \underline{\lambda}(W_{n,x})^{-1}\le\underline{\lambda}(W_x)^{-1}+\varepsilon=\underline{\kappa}^{-1}.$$

Let $\epsilon'=\min\{\delta,\frac{2\underline{\kappa}^2\epsilon}{\overline{\kappa}}\}$. Then, there is $N_{\epsilon',\eta}$ such that $\|W_x-W_{n,x}\|_{op}\le\epsilon'$ so that

$$|h(p,\mathbb{B}_{n,x})^2-h(p,\mathbb{B}_x)^2|\le\underline{\kappa}^{-2}\|W_x-W_{n,x}\|_{op}\le\frac{2\epsilon}{\overline{\kappa}}, \tag{79}$$

with probability at least $1-\eta$ uniformly across P for all $n\ge N_{\epsilon',\eta}$, where the first inequality follows from (78).

Note also that

$$h(p,\mathbb{B}_x)^2=p^\top W_x^{-1}p\ge\underline{\lambda}(W_x^{-1})=\overline{\lambda}(W_x)^{-1}\ge\overline{\kappa}^{-1}. \tag{80}$$

Similarly, again by the Lipschitz continuity of $\overline{\lambda}$ for real symmetric matrices and letting $\varepsilon'=\lambda(W)^{-1}-\overline{\kappa}^{-1}$, there exists $\delta'>0$ such that

$$\|W_x-W_{n,x}\|_{op}\le\delta' \;\Rightarrow\; \overline{\lambda}(W_{n,x})^{-1}\ge\overline{\lambda}(W_x)^{-1}-\varepsilon'=\overline{\kappa}^{-1} \tag{81}$$

implying

$$h(p,\mathbb{B}_{n,x})^2=p^\top W_{n,x}^{-1}p\ge\underline{\lambda}(W_{n,x}^{-1})=\overline{\lambda}(W_{n,x})^{-1}\ge\overline{\kappa}^{-1}. \tag{82}$$

Let $\epsilon''=\min\{\delta',\epsilon'\}$. Noting that $A^2-B^2=(A-B)(A+B)$ and by (79), (80), and (82), there is $N_{\epsilon'',\eta}$ such that

$$|h(p,\mathbb{B}_{n,x})-h(p,\mathbb{B}_x)|=\frac{|h(p,\mathbb{B}_{n,x})^2-h(p,\mathbb{B}_x)^2|}{h(p,\mathbb{B}_{n,x})+h(p,\mathbb{B}_x)}\le\frac{\overline{\kappa}}{2}\times\frac{2\epsilon}{\overline{\kappa}}=\epsilon, \tag{83}$$

with probability at least $1-\eta$ uniformly across P for all $n\ge N_{\epsilon'',\eta}$. Note that the bound in the above expression does not depend on p nor x, and hence it is uniform across $p\in\mathbb{S}^{|Y|-1}$ and $x\in X$. The conclusion of the lemma then follows from the isometry in (74) and letting $\bar{N}_{\epsilon,\eta}=N_{\epsilon'',\eta}$. $\square$

Let $Z=Y\times X$. Let $I_x:Z\times\{0,1\}^{|Y|}$ be defined by $I_x(w)=(1\{y_\ell=y_1,x_\ell=x\},\ldots,1\{y_\ell=y_{|Y|},x_i=$

$x\})'$. Define

$$\mathcal{M}_{P,x} \equiv \Big\{f_{\tilde{b},x} : f_{\tilde{b},x}(z) = \frac{\tilde{b}^\top I_x(z)}{p_x}, \tilde{b} \in \tilde{B}_x\Big\}, \tag{84}$$

and let $\mathcal{M}_P = \bigcup_{x\in X} \mathcal{M}_{P,x}$, where note that X is finite. The following lemma characterizes $\mathcal{M}_P$'s uniform entropy.

Lemma 8. *Suppose Assumptions 4 (ii) and 5 hold. Then, there exist constants $K, v > 0$ that do not depend on P such that*

$$\sup_Q N(\epsilon\|F\|_{L^2_Q}, \mathcal{M}_P, L^2_Q) \leq K\epsilon^{-v},\ \ 0 < \epsilon < 1, \tag{85}$$

where the supremum is taken over all discrete distributions, and F is the envelope function for $\mathcal{M}_P$.

Proof. We first construct an envelope function. Observe that

$$|f_{b,x}(z)| \leq \underline{\zeta}^{-1}|b^\top I_x(z)| \leq \underline{\zeta}^{-1}\|b\|\|I_x(z)\| \leq \underline{\zeta}^{-1}\sqrt{|Y|}\|b\|, \tag{86}$$

where the second inequality is due to the Cauchy-Schwarz inequality. Note that

$$\sup_{b\in\mathbb{B}_x} \|b\|^2 = \sup_{\bar{b}^\top\bar{b}\leq 1}\ \bar{b}^\top W^{-1}\bar{b} \leq \underline{\lambda}(W)^{-1} \leq \kappa^{-1}, \tag{87}$$

where $\bar{b} = W^{1/2}b$. Therefore, one can take $F(z) = \kappa^{-1/2}\underline{\zeta}^{-1}\sqrt{|Y|}$ as the envelope for $\mathcal{M}_P$.

Let $x \in X$ be fixed. Then,

$$|f_{b,x}(z) - f_{b^\top,x}(z)| \leq \underline{\zeta}^{-1}\|I_x(z)\|\|b - b^\top\| \leq C\underline{\zeta}^{-1}\|I_x(z)\|\|b - b^\top\|_{W_x}, \tag{88}$$

for some $C > 0$, where the second inequality follows from the equivalence of norms in a Euclidean space. Note also that $\|I_x(z)\| \leq |Y|^{1/2}$. Following the argument in the proof of Theorem 2 in Andrews (1994), it follows that

$$\sup_Q N(\epsilon\|F\|_{L^2_Q}, \mathcal{M}_{P,x}, L^2_Q) \leq K\epsilon^{-v}, 0 < \epsilon < 1, \tag{89}$$

with $v = |Y|$. Note that $\mathcal{M}_P$ is a finite union of $\mathcal{M}_{P,x}, x \in X$. The conclusion of the lemma then follows by arguing as in the proof of Theorem 3 in Andrews (1994) (see Eq. (A.4)). □

We define the variance semimetric $\rho_P\big((b_1, x_1), (b_2, x_2)\big)$ pointwise by

$$\rho_P\big((b_1, x_1), (b_2, x_2)\big) = \mathrm{Var}\big(f_{b_1,x_1}(z) - f_{b_2,x_2}(z)\big)^{1/2}. \tag{90}$$

Lemma 9. *For any $(b_1, x_1), (b_2, x_2) \in K \times X$, let $\rho\big((b_1, x_1), (b_2, x_2)\big) = \|b_1 - b_2\| + d(x_1, x_2)$, where d is the discrete metric. Let $0 < \delta < 1$. Suppose that $\rho\big((b_1, x_1), (b_2, x_2)\big) \leq \delta$. Then, there exists $C > 0$ such that, for any $P \in \mathcal{P}$,*

$$\rho_P\big((b_1, x_1), (b_2, x_2)\big) \leq C\delta. \tag{91}$$

Proof. Let $0 < \delta < 1$. Suppose that $\rho\big((b_1, x_1), (b_2, x_2)\big) \leq \delta$. Since d is the discrete metric, it must be the case that $\|b_1 - b_2\| \leq \delta$ and $x_1 = x_2 = \bar{x}$ for some $\bar{x} \in X$. By elementary calculation,

$$\mathrm{Var}\big(f_{b_1,x_1}(z) - f_{b_2,x_2}(z)\big) \leq E[|f_{b_1,x_1}(z) - f_{b_2,x_2}(z)|^2] + E[f_{b_1,x_1}(z) - f_{b_2,x_2}(z)]^2 \tag{92}$$

Observe that

$$E[|f_{b_1,x_1}(z) - f_{b_2,x_2}(z)|^2] = E\big[\big((b_1 - b_2)' I_{\bar{x}}(z)\big)^2\big] \leq E[\|b_1 - b_2\|^2 \|I_{\bar{x}}(z)\|^2] \leq |Y|^2 \|b_1 - b_2\|^2. \tag{93}$$

Note also that

$$|E[f_{b_1,x_1}(z) - f_{b_2,x_2}(z)]| = |(b_1 - b_2)' E[I_{\bar{x}}(z)]| \tag{94}$$

$$\leq \|b_1 - b_2\| \|E[I_{\bar{x}}(z)]\| \leq |Y| \|b_1 - b_2\|, \tag{95}$$

where the second inequality is by the Cauchy-Schwarz inequality. Hence, the second term in (92) is bounded by $|Y|^2 \|b_1 - b_2\|^2$. These results imply

$$\mathrm{Var}\big(f_{b_1,x_1}(z) - f_{b_2,x_2}(z)\big)^{1/2} \leq \sqrt{2}|Y| \|b_1 - b_2\| \leq C\delta, \tag{96}$$

where $C = \sqrt{2}|Y|$. The conclusion of the lemma then follows. □

Below, for any $f \in \ell^\infty(K \times X)$, we write Pf to denote its expectation with respect to P. Note that we may write the bootstrapped empirical process $\mathbb{G}_n^*$ as

$$\mathbb{G}_n^*(b, x) = \frac{1}{\sqrt{n}} \sum_{\ell=1}^{n} (M_{n\ell} - 1) \frac{b^\top I_x(z_\ell)}{p_x}, \tag{97}$$

where, for each ℓ, $M_{n\ell}$ is the number of times that $z_\ell = (y_\ell, x_\ell)$ is redrawn from the original sample. We let $P_{M_n}(\cdot | \{y_\ell, x_\ell\}_{\ell=1}^\infty)$ be the conditional law of $M_n = (M_{n1}, \ldots, M_{nn})'$ conditional on the sample path $\{y_\ell, x_\ell\}_{\ell=1}^\infty$ (see van der Vaart and Wellner (1996) Ch. 3.6).

Lemma 10. *Suppose Assumptions 1-3, and 5 hold. Then, for any $\epsilon > 0$ and $\eta > 0$, there exists $\delta > 0$ and $N_{\epsilon,\eta}$ such that*

$$P_{M_n}\Big(\sup_{\rho_P((b_1,x_1),(b_2,x_2)) \leq \delta} \Big|\mathbb{G}_n^*(b_1, x_1) - \mathbb{G}_n^*(b_2, x_2)\Big| \geq \eta \Big| \{y_\ell, x_\ell\}_{\ell=1}^\infty\Big) \leq \epsilon \tag{98}$$

for all $n \geq N_{\epsilon,\eta}$ and for almost all sample paths $\{y_\ell, x_\ell\}_{\ell=1}^\infty$.

Proof. Observe that

$$\mathbb{G}_n^*(b, x) = \sqrt{n}\Big(\frac{1}{n} \sum_{\ell=1}^{n} \frac{b^\top I_x(z_\ell^*)}{p_x} - E[\frac{b^\top I_x(z_\ell)}{p_x}]\Big). \tag{99}$$

Below, we mimic the argument in van der Vaart and Wellner (1996) (Ch.2.5) to show the stochastic equicontinuity of empirical processes. For any $\delta > 0$, define

$$\mathcal{M}_{P,\delta} \equiv \Big\{ f_{b_1,x_1}(z) - f_{b_2,x_2}(z) \ \big| \ \rho_P((b_1,x_1),(b_2,x_2)) \le \tilde{\delta}, b_j \in \mathbb{B}_{x_j}, x_j \in X, j = 1,2 \Big\}. \tag{100}$$

Let $Z_n^*(\delta) = \sup_{f \in \mathcal{M}_{P,\delta}} |\sqrt{n}(\hat{P}_n^* - \hat{P}_n)f|$. Note that by Lemma 9,

$$\rho\big((b_1,x_1),(b_2,x_2)\big) \le \delta \ \Rightarrow \ \rho_P\big((b_1,x_1),(b_2,x_2)\big) \le \tilde{\delta}, \tag{101}$$

for $\tilde{\delta} = C\delta$ with a uniform constant $C > 0$. It then follows that

$$P_{M_n}\Big(\sup_{\rho((b_1,x_1),(b_2,x_2)) \le \delta} \big|\mathbb{G}_n^*(b_1,x_1) - \mathbb{G}_n^*(b_2,x_2)\big| \ge \eta \big| \{y_\ell, x_\ell\}_{\ell=1}^\infty \Big) \le P_{M_n}\Big(Z_n^*(\tilde{\delta}) > \eta \big| \{y_\ell, x_\ell\}_{\ell=1}^\infty \Big). \tag{102}$$

By Markov's inequality and Lemma 2.3.1 in van der Vaart and Wellner (1996), one has

$$\begin{aligned} P_{M_n}\big(Z_n^*(\tilde{\delta}_n) > \eta \big| \{y_\ell, x_\ell\}_{\ell=1}^\infty\big) &\le \frac{2}{\eta} E_{P_{M_n} \times P^e} \left[\sup_{f \in \mathcal{M}_{P,\tilde{\delta}}} \Big| \frac{1}{\sqrt{n}} \sum_{\ell=1}^n e_\ell f(z_\ell^b) \Big| \Big| \{y_\ell, x_\ell\}_{\ell=1}^\infty \right] && (103) \\ &= \frac{2}{\eta} E_{P_{M_n}} \left[E_{P^e} \left[\sup_{f \in \mathcal{M}_{P,\tilde{\delta}}} \Big| \frac{1}{\sqrt{n}} \sum_{\ell=1}^n e_\ell f(z_\ell^b) \Big| \Big| \{z_i^b = \ell\}, \{y_\ell, x_\ell\}_{\ell=1}^\infty \right] \Big| \{y_\ell, x_\ell\}_{\ell=1}^\infty \right], && (104) \end{aligned}$$

where $\{e_\ell\}_{\ell=1}^n$ are i.i.d. Rademacher random variables independent of $\{y_\ell, x_\ell\}_{\ell=1}^\infty$ and $\{M_n\}$. By Hoeffding's inequality, the stochastic process $f \mapsto \{n^{-1/2} \sum_{\ell=1}^n e_i f(z_\ell)\}$ is sub-Gaussian for the $L^2_{\hat{P}_n}$ seminorm $\|f\|_{L^2_{\hat{P}_n}} = (n^{-1} \sum_{\ell=1}^n f(z_\ell)^2)^{1/2}$. By the maximal inequality (Corollary 2.2.8) and arguing as in the proof of Theorem 2.5.2 in in van der Vaart and Wellner (1996), one then has

$$\begin{aligned} E_{P^e} \left[\sup_{f \in \mathcal{M}_{\tilde{\delta}}} \Big| \frac{1}{\sqrt{n}} \sum_{\ell=1}^n e_\ell f(z_\ell^b) \Big| \right] &\le K \int_0^{\tilde{\delta}} \sqrt{\ln N(\epsilon, \mathcal{M}_{P,\tilde{\delta}}, L^2_{\hat{P}_n})} d\epsilon \\ &\le K \|F\|_n \int_0^{\tilde{\delta}/\|F\|_n} \sup_Q \sqrt{\ln N(\epsilon \|F\|_{L^2_Q}, \mathcal{M}_P, L^2_Q)} d\epsilon \\ &\le K' \|F\|_n \int_0^{\tilde{\delta}/\|F\|_n} \sqrt{-v \ln \epsilon} d\epsilon, \end{aligned} \tag{105}$$

for some $K' > 0$, where $\|F\|_n = \sqrt{\frac{1}{n} \sum_{\ell=1} F^2(z_\ell^b)}$, and the last inequality follows from Lemma 8. Note that $\sqrt{-\ln \epsilon} \ge -\ln \epsilon$ for $0 < \epsilon < \tilde{\delta}/\|F\|_n$ with $\tilde{\delta}$ small enough and $\int_0^{\tilde{\delta}/\|F\|_n} - \ln \epsilon d\epsilon = \tilde{\delta}/\|F\|_n (1 - \ln(\tilde{\delta}/\|F\|_n))$. Furthermore, by taking $F(w) = \kappa^{-1/2} \underline{\zeta}^{-1} \sqrt{|Y|}$ (see Proof of Lemma 8), $\|F\|_n = \kappa^{-1/2} \underline{\zeta}^{-1} \sqrt{|Y|} =: \chi > 0$ uniformly across P. In sum, by choosing δ (and hence $\tilde{\delta}$) small enough, one has

$$P_{M_n}\big(Z_n^*(\tilde{\delta}_n) > \eta \big| \{y_\ell, x_\ell\}_{\ell=1}^\infty\big) \le \frac{2\tilde{\delta}}{\eta} (1 - \ln(\tilde{\delta}/\chi)) \le \epsilon, \tag{106}$$

for all n sufficiently large. This establishes the claim of the lemma. □

Lemma 11. *Suppose Assumptions 1-5 hold. For each n, let A_{1n} and A_{2n} be defined as in (49) and (50)* *respectively. Then, for any continuity point of the distribution function of* $\sup_{(b,x)\in\Psi_\infty}\mathbb{G}(b,x)$,

$$P_n\Bigg(\Big\{\big|P_{M_n}\big(\sup_{(b,x)\in\Psi_n}\mathbb{G}_n^*(b,x)\le a\big|y^n,x^n\big)-P\big(\sup_{(b,x)\in\Psi_\infty}\mathbb{G}(b,x)\le a\big)\big|>\epsilon\Big\}\cap A_{1n}^c\cap A_{2n}^c\Bigg)\to 0. \tag{107}$$

as $n\to\infty$.

Proof. Define a metric ρ on $K\times X$ by $\rho((b,x),(b^\top,x'))=\|b-b^\top\|+d(x,x')$, where d is the discrete metric on X. We then let the Hausdorff distance on subsets of $K\times X$ be

$$d_H(A,B)=\max\big\{\sup_{a\in A}\inf_{b\in B}\rho(a,b),\sup_{b\in B}\inf_{a\in A}\rho(a,b)\big\}.$$

For notational simplicity, we use d_H, the same notation as the Hausdorff distance for subsets of K here. Let

$$D_n\equiv\{(b,x):b\in\mathbb{B}_{n,x},x\in X\},\quad\text{and}\quad D\equiv\{(b,x):b\in\mathbb{B}_x,x\in X\}. \tag{108}$$

Suppose that, for any $x\in X$, $d_H(b_{n,x},b_x)<\delta$. Then, by the construction above and $d_H(X,X)=0$ due to X being finite, we have $d_H(D_n,D)=\sup_{x\in X}d_H(b_{n,x},b)+d_H(X,X)<\delta$. Note that Ψ_n and Ψ_∞ can be expressed as subsets of D_n and D as follows

$$\Psi_n=D_n\cap\{(b,x)\in K\times X:\pi^*(b,x)=0\},\quad\text{and}\quad\Psi_\infty=D\cap\{(b,x)\in K\times X:\pi^*(b,x)=0\}. \tag{109}$$

This therefore implies $d_H(\Psi_n,\Psi_\infty)\le d_H(D_n,D)<\delta$.

Let $(b_n^*,x_n^*)\in\text{argmax}_{(b,x)\in\Psi_n}\mathbb{G}_n^*(b,x)$. Let $\Pi_{\Psi_\infty}(b_n^*,x_n^*)$ be the projection of (b_n^*,x_n^*) on Ψ_∞ and note that $\|(b_n^*,x_n^*)-\Pi_{\Psi_\infty}(b_n^*,x_n^*)\|\le d_H(\Psi_n,\Psi_\infty)<\delta$. This implies

$$\begin{aligned}\sup_{(b,x)\in\Psi_n}\mathbb{G}_n^*(b,x)-\sup_{(b,x)\in\Psi_\infty}\mathbb{G}_n^*(b,x)&\le\mathbb{G}_n^*(b_n^*,x_n^*)-\mathbb{G}_n^*(\Pi_{\Psi_\infty}(b_n^*,x_n^*))\\&\le\sup_{\rho((b,x),(b^\top,x'))\le\delta}\big|\mathbb{G}_n^*(b,x)-\mathbb{G}_n^*(b^\top,x')\big|\end{aligned} \tag{110}$$

A similar argument gives

$$\sup_{(b,x)\in\Psi_\infty}\mathbb{G}_n^*(b,x)-\sup_{(b,x)\in\Psi_n}\mathbb{G}_n^*(b,x)\le\sup_{\rho((b,x),(b^\top,x'))\le\delta}\big|\mathbb{G}_n^*(b,x)-\mathbb{G}_n^*(b^\top,x')\big|. \tag{111}$$

Hence, for any $\eta>0$,

$$\big|\sup_{(b,x)\in\Psi_n}\mathbb{G}_n^*(b,x)-\sup_{(b,x)\in\Psi_\infty}\mathbb{G}_n^*(b,x)\big|\ge\eta\ \Rightarrow\ \sup_{\rho((b,x),(b^\top,x'))\le\delta}\big|\mathbb{G}_n^*(b,x)-\mathbb{G}_n^*(b^\top,x')\big|\ge\eta. \tag{112}$$

Now suppose $(y^n,x^n)\in A_{1n}^c\cap A_{2n}^c$, where A_{1n} and A_{2n} are defined as in (49) and (50) respectively. Then, for any $\eta > 0$, there is $\delta > 0$ such that

$$\begin{aligned}
&P_{M_n}\big(\sup_{(b,x)\in\Psi_n} \mathbb{G}_n^*(b,x) \le a\big|y^n, x^n\big) \\
&\quad\le P\Big(\big|\sup_{(b,x)\in\Psi_n} \mathbb{G}_n^*(b,x) - \sup_{(b,x)\in\Psi_\infty} \mathbb{G}_n^*(b,x)\big| \ge \eta\big|y^n, x^n\Big) + P_{M_n}\big(\sup_{(b,x)\in\Psi_\infty} \mathbb{G}_n^*(b,x) \le a+\eta\big|y^n, x^n\big) \\
&\quad\le P_{M_n}\big(\sup_{\rho((b,x),(b^\top,x'))\le\delta} \big|\mathbb{G}_n^*(b,x) - \mathbb{G}_n^*(b^\top,x')\big| \ge \eta\big|y^n, x^n\big) + P_{M_n}\big(\sup_{(b,x)\in\Psi_\infty} \mathbb{G}_n^*(b,x) \le a+\eta\big|y^n, x^n\big) \\
&\quad\le \frac{\epsilon}{3} + P_{M_n}\big(\sup_{(b,x)\in\Psi_\infty} \mathbb{G}_n^*(b,x) \le a+\eta\big|y^n, x^n\big), \qquad (113)
\end{aligned}$$

for all n sufficiently large, where the last inequality follows from (112) and Lemma 10. Therefore, by (113) and the triangle inequality,

$$\begin{aligned}
&\Big|P_{M_n}\big(\sup_{(b,x)\in\Psi_n} \mathbb{G}_n^*(b,x) \le a\big|y^n, x^n\big) - P\big(\sup_{(b,x)\in\Psi_\infty} \mathbb{G}(b,x) \le a\big)\Big| \\
&\qquad\le \frac{\epsilon}{3} + \Big|P_{M_n}\big(\sup_{(b,x)\in\Psi_\infty} \mathbb{G}_n^*(b,x) \le a+\eta\big|y^n, x^n\big) - P\big(\sup_{(b,x)\in\Psi_\infty} \mathbb{G}(b,x) \le a+\eta\big)\Big| \\
&\qquad\quad + \Big|P\big(\sup_{(b,x)\in\Psi_\infty} \mathbb{G}(b,x) \le a+\eta\big) - P\big(\sup_{(b,x)\in\Psi_\infty} \mathbb{G}(b,x) \le a\big)\Big| \le \epsilon, \qquad (114)
\end{aligned}$$

for any $\eta > 0$ and for all n sufficiently large, where the last inequality follows from

$$\Big|P_{M_n}\big(\sup_{(b,x)\in\Psi_\infty} \mathbb{G}_n^*(b,x) \le a+\eta\big|y^n, x^n\big) - P\big(\sup_{(b,x)\in\Psi_\infty} \mathbb{G}(b,x) \le a+\eta\big)\Big| \le \frac{\epsilon}{3}, \qquad (115)$$

for all n sufficiently large by Theorem 3.6.2 in van der Vaart and Wellner (1996) and the portmanteau theorem. Finally, since a is a continuity point, one can choose η sufficinetly small sothat

$$\Big|P\big(\sup_{(b,x)\in\Psi_\infty} \mathbb{G}(b,x) \le a+\eta\big) - P\big(\sup_{(b,x)\in\Psi_\infty} \mathbb{G}(b,x) \le a\big)\Big| \le \frac{\epsilon}{3}. \qquad (116)$$

Hence, (114) establishes the claim of the lemma. □

B.4 Results on Multiple Hypothesis Testing

Proof of Theorem 3. Let $\{H_j : j \in \mathcal{J}\}$ be the set of true null hypotheses with $\mathcal{J} \subseteq \{1, ..., J\}$. Let k be the smallest index satisfying

$$p_{(k)} = \min_{j\in\mathcal{J}} p_j.$$

Let $j^* \equiv \arg\min_{j\in\mathcal{J}} p_j$. Then, $p_{(k)} = p_{j^*}$. Note that if H_{j*} is true, then H_j is logically true $\forall j > j^*$, because $S \succeq S^r_{j^*}$ implies $S \succeq S^r_j$ $\forall j > j^*$. Therefore, $\{1, ..., j^* - 1\}$ is the set of nulls that remain to be tested. Let $\mathcal{J}^* \equiv \mathcal{J} \cap \{1, ..., j^*\}$ be the set of true nulls among $\{1, ..., j^*\}$. Note that

$$\min_{j\in\mathcal{J}} p_j = \min_{j\in\mathcal{J}^*} p_j.$$

Also note that, by construction, there are at most $j^* - (k-1)$ true nulls among $\{1, ..., j^*\}$ and so $|\mathcal{J}^*| \leq j^* - (k-1)$. This is because of the following argument. Recall that k is the first step that a true null is being tested; in other words, there should be at least $k-1$ nulls that are false. Because $p_{(k)} = p_{j^*}$ and H_j is true $\forall j \geq j^*$, those false nulls should be included in $\{1, ..., j^* - 1\}$. Therefore, $\{1, ..., j^* - 1\}$ should contain at most $j^* - 1 - (k-1)$ true nulls.

Now, Algorithm 1 commits a false rejection if

$$p_{(1)} \leq c_1, ..., p_{(k)} \leq c_k,$$

which trivially implies

$$\min_{j \in \mathcal{J}^*} p_j = p_{(k)} \leq c_k = \frac{\alpha}{j^* - k + 1} \leq \frac{\alpha}{|\mathcal{J}^*|}.$$

Therefore, for given P and n and conditional on j^*, we have the probability of a false rejection, $E_P[\phi^{MH}]$, bounded above by

$$P\left[\min_{j \in \mathcal{J}^*} p_j \leq \alpha / |\mathcal{J}^*|\right] = P\left[\bigcup_{j \in \mathcal{J}^*} \{p_j \leq \alpha / |\mathcal{J}^*|\}\right] \leq \sum_{j \in \mathcal{J}^*} P\left[p_j \leq \alpha / |\mathcal{J}^*|\right],$$

where the inequality is the Bonferroni inequality. Finally, conditional on j^*,

$$\limsup_{n \to \infty} \sup_{P \in \mathcal{P}} E_P[\phi^{MH}] \leq \sum_{j \in \mathcal{J}^*} \limsup_{n \to \infty} \sup_{P \in \mathcal{P}} P\left[p_j \leq \alpha / |\mathcal{J}^*|\right] \leq \alpha,$$

where the second inequality is by Theorem 2. By the law of iterated expectation (applied before taking the lim sup and sup), the above result will hold without conditioning on j^*. $\square$

C Details on the Monte Carlo Experiments

C.1 BNE Threshold

Recall that $\Psi(\tau) = P(\varepsilon_i \geq \tau)$. The threshold in the equilibrium strategy in (23) solves

$$\tau_i(\nu_i, t_i) = -(x'\beta + \Delta E_{\nu_{-i}, t_{-i}}[\Psi(\tau_{-i}(\nu_{-i}, t_{-i})) \mid \nu_i, t_i] + \nu_i), \tag{117}$$

where

$$E_{\nu_{-i}, t_{-i}}[\Psi(\tau_{-i}(\nu_{-i}, t_{-i})) \mid t_i, \nu_i] = q E_{\nu_{-i}}[\Psi(\tau_{-i}(\nu_{-i}, \nu_i)) \mid t_i] + (1-q) E_{\nu_{-i}}[\Psi(\tau_{-i}(\nu_{-i}, -\nu_i)) \mid t_i] \tag{118}$$

Note moreover that

$$E_{\nu_{-i}}[\Psi(\tau_{-i}(\nu_{-i}, \nu_i)) \mid t_i = \eta] = \rho_\eta(p, q) \Psi(\tau_{-i}(\eta, \nu_i)) + (1 - \rho_\eta(p, q)) \Psi(\tau_{-i}(-\eta, \nu_i)) \tag{119}$$

$$E_{\nu_{-i}}[\Psi(\tau_{-i}(\nu_{-i}, \nu_i)) \mid t_i = -\eta] = \rho_{-\eta}(p, q) \Psi(\tau_{-i}(-\eta, \nu_i)) + (1 - \rho_{-\eta}(p, q)) \Psi(\tau_{-i}(\eta, \nu_i)) \tag{120}$$

so that we can create a system of equations to solve simultaneously for all thresholds $\tau_i(\nu_i, t_i)$. In fact, we can write:

$$
\begin{aligned}
\tau_i(\nu_i, \eta) &= -(x'\beta + \Delta q \left[\rho_\eta(p,q) P(\tau_{-i}(\eta, \nu_i)) + (1-\rho_\eta(p,q))\Psi(\tau_{-i}(-\eta, \nu_i))\right] + \\
&\quad + \Delta(1-q) \left[\rho_\eta(p,q)\Psi(\tau_{-i}(\eta, -\nu_i)) + (1-\rho_\eta(p,q))\Psi(\tau_{-i}(-\eta, -\nu_i))\right]] + \nu_i) \\
\tau_i(\nu_i, -\eta) &= -(x'\beta + \Delta q \left[\rho_{-\eta}(p,q)\Psi(\tau_{-i}(-\eta, \nu_i)) + (1-\rho_{-\eta}(p,q))\Psi(\tau_{-i}(\eta, \nu_i))\right] + \\
&\quad + \Delta(1-q) \left[\rho_{-\eta}(p,q)\Psi(\tau_{-i}(-\eta, -\nu_i)) + (1-\rho_{-\eta}(p,q))\Psi(\tau_{-i}(\eta, -\nu_i))\right]] + \nu_i)
\end{aligned}
$$

and the system to solve for a symmetric equilibrium, i.e., an equilibrium where thresholds are equal across firms, is:

$$
\begin{aligned}
\tau(\eta, \eta) &= -(x'\beta + \Delta q \left[\rho_\eta(p,q)\Psi(\tau(\eta, \eta)) + (1-\rho_\eta(p,q))\Psi(\tau(-\eta, \eta))\right] + \\
&\quad + \Delta(1-q) \left[\rho_\eta(p,q)\Psi(\tau(\eta, -\eta)) + (1-\rho_\eta(p,q))\Psi(\tau(-\eta, -\eta))\right]] + \eta) \\
\tau(-\eta, \eta) &= -(x'\beta + \Delta q \left[\rho_\eta(p,q)\Psi(\tau(\eta, -\eta)) + (1-\rho_\eta(p,q))\Psi(\tau(-\eta, -\eta))\right] + \\
&\quad + \Delta(1-q) \left[\rho_\eta(p,q)\Psi(\tau(\eta, \eta)) + (1-\rho_\eta(p,q))\Psi(\tau(-\eta, \eta))\right]] - \eta) \\
\tau(\eta, -\eta) &= -(x'\beta + \Delta q \left[\rho_{-\eta}(p,q)\Psi(\tau(-\eta, \eta)) + (1-\rho_{-\eta}(p,q))\Psi(\tau(\eta, \eta))\right] + \\
&\quad + \Delta(1-q) \left[\rho_{-\eta}(p,q)\Psi(\tau(-\eta, -\eta)) + (1-\rho_{-\eta}(p,q))\Psi(\tau(\eta, -\eta))\right]] + \eta) \\
\tau(-\eta, -\eta) &= -(x'\beta + \Delta q \left[\rho_{-\eta}(p,q)\Psi(\tau(-\eta, -\eta)) + (1-\rho_{-\eta}(p,q))\Psi(\tau(\eta, -\eta))\right] + \\
&\quad + \Delta(1-q) \left[\rho_{-\eta}(p,q)\Psi(\tau(-\eta, \eta)) + (1-\rho_{-\eta}(p,q))\Psi(\tau(\eta, \eta))\right]] - \eta).
\end{aligned}
$$