Identification and inference in discrete choice models with imperfect information*

Cristina Gualdani[†] Shruti Sinha[‡]

January 2020

Abstract

We study identification and inference of preference parameters in a single-agent, static, discrete choice model where the decision maker may face attentional limits precluding her to exhaustively acquire information about the payoffs of the available alternatives. Instead of explicitly modelling the information constraints, which can be susceptible to misspecification, we leverage on the notion of one-player Bayesian Correlated Equilibrium in Bergemann and Morris (2016) to provide a tractable characterisation of the sharp identified set and discuss inference under minimal assumptions on the amount of information processed by the decision maker. Simulations reveal that the obtained bounds on the preference parameters can be tight in several settings of empirical interest.

Keywords: Discrete choice model, Bayesian Persuasion, Bayes Correlated Equilibrium, Incomplete Information, Partial Identification, Moment Inequalities.

*We acknowledge funding from the French National Research Agency (ANR) under the Investments for the Future (Investissements d'Avenir) program, grant ANR-17-EURE-0010.

†Email: cristina.gualdani@tse-fr.eu, Toulouse School of Economics, University of Toulouse Capitole, Toulouse, France.

‡Email: shruti.sinha@tse-fr.eu, Toulouse School of Economics, University of Toulouse Capitole, Toulouse, France.

1 Introduction

Attentional limits have long been recognised to play a critical role in decision problems because they preclude agents' ability to exhaustively process information about the values of the available alternatives (e.g., Simon, 1955; 1959; Kaheman, 1973; Sims, 1998; 2003; 2006; Lacetera, Pope, and Sydnor, 2012; Des Los Santos, Hortaçsu, and Wildenbeest, 2012; Matějka and McKay, 2015; Caplin, Dean, and Leahy, 2019b). In this paper we offer a robust and tractable method to incorporate attentional limits in the empirical analysis of decision problems. In particular, we study identification and inference of preferences in a single-agent, static, discrete choice model where the decision maker (hereafter, DM) may face attentional limits hampering her capacity to learn about the payoffs generated by the available alternatives.

More formally, we consider a static setting where the DM has to choose an alternative from a finite set. The payoff generated by each alternative depends on the state of the world.[1] The DM chooses an alternative possibly without being fully aware of the state of the world. Instead, the DM has a prior on the state of the world. Moreover, the DM has the opportunity to investigate further the state of the world by processing additional information (hereafter, *information structure*, as per Bergemann and Morris, 2013; 2016). This information structure takes the form of a noisy signal of the state of the world and can range from full revelation of the state of the world to no information whatsoever, depending on the DM's attentional limits. The DM uses the acquired information structure to update her prior and obtain a posterior. Finally, the DM chooses an alternative maximising the expected payoff, where the expectation is computed via the posterior.

We assume that the researcher has data on choices made by many i.i.d. DMs facing the decision problem above and, possibly, one some covariates which are part of (or coincide with) the state of the world. However, the researcher does not observe information structures, which remain latent, potentially heterogenous across agents, and arbitrarily correlated with the payoff-relevant variables. Our objective is studying identification and inference of the preference parameters (specifically, payoff functions and distributions of unobservables) from the empirical choice probabilities. In doing so, we want to remain *agnostic about information structures* and, thus, allow agents to compute expected payoffs with *any Bayes-plausible posterior*.

We believe that performing this exercise is useful to answer relevant questions. First, we can use our methodology to robustly investigate the stability of the preference parameters across decision contexts. That is, we can test if changes in substitution patterns across decision contexts are only due to changes in the distribution of covariates and/or in information structures, or they can also be due to changes in the preference parameters. Second, our methodology allows to flexibly incorporate any minimal level of information available to agents before choosing. Hence, we can use our methodology to perform a sensitivity analysis of the

[1]The state of the world is defined by the realisation of variables like attributes of the available alternatives, attributes and tastes of the DM, exogenous market shocks, etc.

empirical conclusions to changes in such minimal level. Third, we can use our methodology to obtain counterfactual choice probabilities when the researcher modifies some payoff-relevant variables and keeps (or not) information structures fixed (Bergemann, Brooks, and Morris, 2019). Fourth, consider the case where the analyst knows more than agents about the state of the world, while agents might only partially learn about it. For instance, in transport choice problems, even the most conscientious consumers are typically unaware of the exact pollution costs of the available transport modes. In health contexts, even the most scrupulous patients are often unable to precisely evaluate the opportunity of drug treatments versus surgery treatments. When selecting an insurance plan, even the most meticulous beneficiaries may not know how to accurately compute expected out-of-pocket costs for each insurance plan. In such settings, policy makers are interested in evaluating the impact on choices of sending agents information about the state of the world (e.g., Hastings, Tejeda-Ashton, 2008; Bettinger, et al., 2012; Kling, et al., 2012). We can use our methodology to address this question by constructing counterfactual choice probabilities when the researcher sends agents information about the state of the world and keeps the payoff-relevant variables fixed. Lastly, note leaving information structures completely unrestricted means that our results are solid against various assumptions on agents' cognitive skills. In fact, the developed framework nests several discrete choice models that have been analysed in the literature, e.g., Logit model, Nested Logit model, Mixed Logit model, discrete choice models with rational inattention (e.g., Caplin and Dean, 2015; Matějka and McKay, 2015), some discrete choice models with search (e.g., Hébert and Woodford, 2018; Morris and Strack, 2019), and discrete choice models with risk aversion (see Barseghyan, Molinari, O'Donoghue, and Teitelbaum, 2018 for a review).

Studying identification and inference of the preference parameters while remaining agnostic about information structures is challenging because the model is incomplete in the sense of Tamer (2003), thus raising the possibility of partially identified preference parameters. Tractably characterising the sharp identified set is not an easy task. In fact, in order to determine whether a given parameter value belongs to the sharp identified set, we need to establish whether the empirical choice probabilities belong to the collection of choice probabilities predicted by our model under a range of possible information structures. The difficulty here lies in the necessity of exploring such a range of possible information structures.

We approach the above problem by applying the notion of one-player Bayes Correlated Equilibrium provided in Bergemann and Morris (2013; 2016).[2] Specifically, we exploit Theorem 1 in Bergemann and Morris (2016) to claim that the collection of choice probabilities predicted by our model under a range of possible information structures is equivalent to the collection of choice probabilities predicted by our model under the notion of one-player Bayes Correlated Equilibrium. That is, it is equivalent to the collection of choice probabilities in

[2]Note that the notion of one-player Bayes Correlated Equilibrium can be related to Bayesian Persuasion (Kamenica and Gentzkow, 2011). More details on this are in footnote 12. Further, Caplin and Martin (2015) discusses the relation between one-player Bayes Correlated Equilibrium, Bayesian Persuasion, and Bayesian Expected Utility maximisation.

a mediated decision problem where the mediator directly gives recommendations to the DMs and these recommendations are incentive compatible. Further, such a collection is a convex set. Therefore, determining whether a given parameter value belongs to the sharp identified set amounts to finding whether the empirical choice probabilities belong to that convex set. By using insights from Magnolfi and Roncoroni (2017) and Syrgkanis, Tamer, and Ziani (2018), we argue that this corresponds to solving a linear programming problem. Thus, constructing the sharp identified set becomes a computationally tractable exercise. Lastly, after having reformulated the identifying restrictions as moment inequalities, we explain how inference on the sharp identified set can be conducted by using Andrews and Shi (2013)'s generalised moment selection procedure. Simulations of discrete choice models with risk aversion and Nested Logit models highlight that our methodology can produce informative bounds for the preference parameters.

Literature review Research questions similar to ours have been addressed in the literature using different approaches. Caplin and Martin (2015) study a related problem by using data on choices for every possible realisation of the state of the world. Instead, in our framework the researcher may fully/partially or may not observe the state of the world. Further, we incorporate more degrees of heterogeneity across agents.

Another approach consists in modelling the mechanism according to which agents acquire information structures. By doing so, the analyst may obtain point identification of the preference parameters, however at the cost of potential misspecification of information constraints. For example, Mehta, Rajiv, and Srinivasan (2003), Honka and Chintagunta (2016), and Abaluck and Compiani (2019) consider search frameworks, where agents follow a protocol to learn about payoffs. Csaba (2018) adopts a rational inattention perspective, where the attentional costs sustained by agents to acquire information structures are parametrically modelled, along the lines of Caplin and Dean (2015), Matějka and McKay (2015), Fosgerau, Melo, de Palma, and Shum (2017), and Caplin, Dean, and Leahy (2019b).

Further, this paper broadly relates to the econometric literature on discrete choice models when the sets of alternatives actually considered by agents (hereafter, *consideration sets*) could be subsets of the entire set of alternatives, heterogenous, arbitrarily correlated with the payoff-relevant variables, and latent (for some recent contributions see, e.g., Abaluck and Adams, 2018; Barseghyan, Coughlin, Molinari, and Teitelbaum, 2019; Barseghyan, Molinari, and Thirkettle, 2019; Cattaneo, Ma, Masatlioglu, and Suleymanov, 2019). In fact, one key implication of attentional limits is that, since attention is a scarce resource, agents may process information structures inducing them to contemplate, in equilibrium, only a subset of the available alternatives, ignoring all the others. Hence, in our model, consideration sets can arise endogenously (Caplin, Dean, and Leahy, 2019b).[3,4] Yet, there is an important difference between the literature on the econometrics of consideration sets and this paper. The former focuses on recovering consideration probabilities from the empirical choice probabilities, but parameterises indirect utilities. The latter allows indirect utilities to depend on any Bayes-plausible posterior, but does not recover consideration probabilities. Thus, we can answer different questions. In particular, we can use our methodology to robustly study stability of preferences across decision contexts, counterfactual predictions when changing payoffs or information structures, and sensitivity analysis of empirical conclusions to agents' minimal level of information.

This paper also relates to the literature concerned with evaluating the impact on choices of sending agents information about the state of the world (e.g., Hastings and Tejeda-Ashton, 2008, studying retirement fund options in Mexico; Bettinger, et al., 2012, studying application to colleges; Kling, et al., 2012, studying Medicare Part D prescription drug plans in the United States). This literature typically proceeds under strong assumptions on how agents process information and/or exploits randomised field experiments. Instead, under very weak restrictions, our methodology can be used to obtain counterfactual choice probabilities when the researcher send agents information about the state of the world and keeps the payoff-relevant variables fixed.

More generally, this work relates to the literature concerned with relaxing assumptions about expectation formation and about the amount of information on which agents condition their expectations (see, e.g., the seminal paper by Manski, 2004). Our paper can be interpreted as a robustness exercise in that direction. In fact, by not restricting information structures, we allow agents to compute expectations with any Bayes-plausible posterior.

Lastly, this paper relates to two important works, Magnolfi and Roncoroni (2017) and Syrgkanis, Tamer, and Ziani (2018), that exploit Theorem 1 in Bergemann and Morris (2016) to characterise the sharp identified set in an entry game framework and in an auction framework, respectively, under latent information structures. Despite relying on a similar technology, our framework is not nested in theirs because, first, we consider a multinomial choice setting. Second, we impose different assumptions on agents' minimal amount of information. In particular, in our framework agents are uncertain about their own payoffs, while in Magnolfi and Roncoroni (2017) and Syrgkanis, Tamer, and Ziani (2018) agents are uncertain about others' payoffs. We contribute to this thread of the literature by highlighting the empirical usefulness of the notion of Bayes Correlated Equilibrium in a single-agent, static, discrete choice model

[3] Recall that the DM's information structure takes the form of a noisy signal of the state of the world. Hence, an alternative belongs to the DM's consideration set if the subset of the signal's support inducing the DM to choose that alternative has positive measure (Caplin, Dean, and Leahy, 2019b). More details are in Section 2.

[4] Limited attention in choice is not the only mechanism that can induce endogenous considerations sets in discrete choice models. Consideration sets may arise also because of lack of awareness of some alternatives in the feasible set (e.g., Goeree, 2008), deliberately ignoring some alternatives in the feasible set (e.g., Wilson, 2008), incomplete product availability (e.g., Conlon and Mortimer, 2014), being offered the possibility of receiving program access from outside an experiment (e.g., Kamat, 2019), and absence of market clearing transfers in two-sided matching models (e.g., He, Sinha, and Sun, 2019).

with attentional limits.

The remainder of the paper is organised as follows. Section 2 describes the model. Section 3 discusses identification. Section 4 presents simulations. Section 5 illustrates inference. Section 6 concludes.

Notation Capital letters are used for random variables/vectors/matrices and small case letters for their realisations. Calligraphic capital letters are used for sets. Given a random vector Z, P_Z denotes its joint density when all the components of Z are continuous, mixed joint density when some components of Z are continuous and some discrete, and probability mass function when all the components of Z are discrete. However, for readability, sometimes in the paper we generically refer to P_Z as a *density*.

$\mathbb{R}^K_+$ denotes the K-dimensional positive real space. Given set $\mathcal{A}$, $\Delta(\mathcal{A})$ is the function space of all possible densities with support equal to or contained in $\mathcal{A}$. Given set $\mathcal{A}$, $|\mathcal{A}|$ denotes $\mathcal{A}$'s cardinality. Given two sets, $\mathcal{A}$ and $\mathcal{R} \subseteq \mathcal{A}$, $\mathcal{A} \setminus \mathcal{R}$ is the complement of $\mathcal{R}$ in $\mathcal{A}$. 0_L is the $L \times 1$ vector of zeros.

Consider two random variables, Z and X, with supports $\mathcal{Z}$ and $\mathcal{X}$, respectively. Given $x \in \mathcal{X}$, we denote the density of Z conditional on $X = x$ by $P_{Z|X}(\cdot|x)$. Further, we denote the family of densities of Z conditional on every realisation x of X by $\mathcal{P}_{Z|X}$, i.e., $\mathcal{P}_{Z|X} \equiv \{P_{Z|X}(\cdot|x) \in \Delta(\mathcal{Z}) : x \in \mathcal{X}\}$. Note that $\mathcal{P}_{Z|X}$ contains $|\mathcal{X}|$ densities.

$\mathbb{S}^{|\mathcal{Y}|}$ is the unit sphere in $\mathbb{R}^{|\mathcal{Y}|}$, i.e., $\mathbb{S}^{|\mathcal{Y}|} \equiv \{b \in \mathbb{R}^{|\mathcal{Y}|} : b^T b = 1\}$. $\mathbb{B}^{|\mathcal{Y}|}$ is the unit ball in $\mathbb{R}^{|\mathcal{Y}|}$, i.e., $\mathbb{S}^{|\mathcal{Y}|} \equiv \{b \in \mathbb{R}^{|\mathcal{Y}|} : b^T b \leq 1\}$.

"$\times$" denotes the Cartesian product operator or is used to indicate vector dimensions. "$\cdot$" denotes the standard product operator.

2 The model

In this section we describe a single-agent, static, discrete choice model, where DM i may be partially informed about the payoffs generated by the available alternatives. Then, we characterise DM i's optimal strategy and discuss some examples.

Let DM i face the decision problem of choosing an alternative from the finite set $\mathcal{Y}$, possibly under incomplete information about the state of the world. The state of the world affects the payoff that DM i gets from the decision problem and is represented by a vector, (x_i, e_i, v_i). x_i is a realisation of some covariates, X_i, with support $\mathcal{X}$ and density P_X. x_i is observed by DM i and the researcher. e_i is a realisation of some tastes of DM i, ϵ_i, with support $\mathcal{E}$ and density conditional on $X_i = x_i$ denoted by $P_{\epsilon|X}(\cdot|x_i) \in \mathcal{P}_{\epsilon|X}$. e_i is observed by DM i but not by the researcher. v_i is a realisation of some further (dis)value, V_i, that DM i can derive from the choice problem, with support $\mathcal{V}$. v_i may or may not be observed by the researcher. v_i is not observed by DM i. However, DM i has a prior on V_i conditional on $(X_i, \epsilon_i) = (x_i, e_i)$, denoted by

$P_{V|X,\epsilon}(\cdot|x_i,e_i) \in \mathcal{P}_{V|X,\epsilon}$.[5,6] Moreover, DM i can refine such a prior upon reception of a private signal, t_i, which may or may not be informative about v_i. In particular, t_i is a realisation of the random variable (or, vector/matrix) T_i, with support $\mathcal{T}_i$ and density conditional on $(X_i,\epsilon_i,V_i) = (x_i,e_i,v_i)$ denoted by $P^i_{T|X,\epsilon,V}(\cdot|x_i,e_i,v_i) \in \mathcal{P}^i_{T|X,\epsilon,V}$. DM i observes t_i, uses $P^i_{T|X,\epsilon,V}(\cdot|x_i,e_i,v_i)$ to update $P_{V|X,\epsilon}(\cdot|x_i,e_i)$, and obtains the posterior, $P^i_{V|X,\epsilon,T}(\cdot|x_i,e_i,t_i) \in \mathcal{P}^i_{V|X,\epsilon,T}$, via the Bayes rule. Finally, DM i chooses alternative $y \in \mathcal{Y}$ maximising her expected payoff, $\int_{v\in\mathcal{V}} u(y,x_i,e_i,v)P^i_{V|X,\epsilon,T}(v|x_i,e_i,t_i)dv$, where $u : \mathcal{Y}\times\mathcal{X}\times\mathcal{E}\times\mathcal{V} \to \mathbb{R}$ is the payoff function. If there is more than one maximising alternative (i.e., if there are ties), then DM i applies some tie-breaking rule.

Before proceeding, note that we distinguish among X_i, ϵ_i, and V_i in order to get a flexible framework nesting various settings. However, the researcher can easily get rid of any among X_i, ϵ_i, and V_i by assuming degenerate supports. Therefore, the researcher has the freedom to decide which is the minimal set of variables observed by DM i before choosing. In some scenarios, the researcher may prefer to be very cautious and assume that none of the variables of the model are certainly observed by DM i before choosing. In other scenarios, the researcher may feel confident of imposing that DM i certainly observes some variables of the model. This choice could have an impact on the identifying power of the model. As explained later, our methodology can also be used to perform a sensitivity analysis of the identifying power of the model to changes in the minimal set of variables observed by agents before choosing.

It is useful to summarise the framework above as follows. $G \equiv (\mathcal{Y},\mathcal{X},\mathcal{E},\mathcal{V},u,\mathcal{P}_{V|X,\epsilon},\mathcal{P}_{\epsilon|X})$ will be hereafter called "*baseline choice problem*". G contains the primitives that the researcher wants to identify, i.e., u, $\mathcal{P}_{V|X,\epsilon}$, and $\mathcal{P}_{\epsilon|X}$. G represents the *minimal* information available to DM i before choosing, together with (x_i,e_i). $S_i \equiv (\mathcal{T}_i,\mathcal{P}^i_{T|X,\epsilon,V})$ will be hereafter called DM i's "*information structure*". S_i represents the *additional* information that DM i processes to form a posterior on V_i, together with t_i. S_i can range from complete revelation of V_i (hereafter, complete information structure[7]) to no information whatsoever on V_i (hereafter, degenerate information structure[8]), depending on DM i's attentional limits. Lastly, the pair (G,S_i) constitutes what will be hereafter called DM i's "*augmented choice problem*".

Note that information structures can vary across agents. This is because different agents could have different attention constraints and, consequently, sustain different costs to gather information on the state of the world. More precisely, even if two agents i,j are such that $(x_i,e_i,v_i) = (x_j,e_j,v_j)$, it may be that $P^i_{T|X,\epsilon,V}(\cdot|x_i,e_i,v_i) \neq P^j_{T|X,\epsilon,V}(\cdot|x_j,e_j,v_j)$ and, hence,

[5] X_i, ϵ_i, and V_i can be scalars, vectors, or matrices. Further, they can be individual-specific, alternative-specific, and pair-specific.

[6] Note that the state of the world can vary across agents, as highlighted by subscript i in (x_i,e_i,v_i). Further, if two agents i,j are such that $(x_i,e_i) \neq (x_j,e_j)$, then $P_{\epsilon|X}(\cdot|x_i)$ and $P_{V|X,\epsilon}(\cdot|x_i,e_i)$ can be different from $P_{\epsilon|X}(\cdot|x_j)$ and $P_{V|X,\epsilon}(\cdot|x_j,e_j)$, respectively.

[7] One representation of the complete information structure is $\mathcal{T}_i \equiv \mathcal{V}, P^i_{T|X,\epsilon,V}(v|x,e,v) = 1\ \forall x \in \mathcal{X}, \forall e \in \mathcal{E}$, and $\forall v \in \mathcal{V}$.

[8] One representation of the degenerate information structure is $\mathcal{T}_i \equiv \{t\}, P^i_{T|X,\epsilon,V}(t|x,e,v) = 1\ \forall x \in \mathcal{X}, \forall e \in \mathcal{E}$, and $\forall v \in \mathcal{V}$, where t is any real number. Note that under the degenerate information structure DM i's posterior is equal to DM i's prior about V_i.

$P^i_{V|X,\epsilon,T}(\cdot|x_i, e_i, t_i) \neq P^j_{V|X,\epsilon,T}(\cdot|x_j, e_j, t_j)$. Also, it may be that $t_i = t_j$ but $P^i_{T|X,\epsilon,V}(\cdot|x_i, e_i, v_i) \neq P^j_{T|X,\epsilon,V}(\cdot|x_j, e_j, v_j)$. Such heterogeneity is highlighted by the superscript/subscript i in $\mathcal{P}^i_{T|X,\epsilon,V}$, $\mathcal{P}^i_{V|X,\epsilon,T}$, $\mathcal{T}_i$, and S_i. Further, note that conditional signal densities can be arbitrarily correlated with agents' preferences. Lastly, we anticipate that in the data generating process (Assumption 1 below), S_i will be treated as unobserved by the researcher and left completely unrestricted. This means that DM i computes her expected payoff with some unspecified and potentially endogenous Bayes-plausible posterior.

We now define DM i's optimal strategy when facing the augmented choice problem (G, S_i). Let Y_i be a random variable representing DM i's choice. A (mixed[9]) strategy in the augmented choice problem (G, S_i) is a family of probability mass functions of Y_i conditional on every realisation (x, e, t) of (X_i, ϵ_i, T_i), i.e.,

$$\mathcal{P}^i_{Y|X,\epsilon,T} \equiv \{P^i_{Y|X,\epsilon,T}(\cdot|x, e, t) \in \Delta(\mathcal{Y}) : x \in \mathcal{X}, e \in \mathcal{E}, t \in \mathcal{T}_i\}.$$

$\mathcal{P}^i_{Y|X,\epsilon,T}$ is an optimal strategy of the augmented choice problem (G, S_i) if, for each $x \in \mathcal{X}$, $e \in \mathcal{E}$, and $t \in \mathcal{T}_i$, DM i maximises her expected payoff by choosing alternative $y \in \mathcal{Y}$ such that $P^i_{Y|X,\epsilon,T}(y|x, e, t) > 0$.

Definition 1. (*Optimal strategy of the augmented choice problem* (G, S_i)) $\mathcal{P}^i_{Y|X,\epsilon,T}$ is an optimal strategy of the augmented choice problem (G, S_i) if $\forall x \in \mathcal{X}$, $\forall e \in \mathcal{E}$, and $\forall t \in \mathcal{T}_i$,

$$\int_{v\in\mathcal{V}} u(y, x, e, v) P^i_{T|X,\epsilon,V}(t|x, e, v) P_{V|X,\epsilon}(v|x, e) dv \geq \int_{v\in\mathcal{V}} u(\tilde{y}, x, e, v) P^i_{T|X,\epsilon,V}(t|x, e, v) P_{V|X,\epsilon}(v|x, e) dv,$$

$\forall y \in \mathcal{Y}$ such that $P^i_{Y|X,\epsilon,T}(y|x, e, t) > 0$, and $\forall \tilde{y} \in \mathcal{Y} \setminus \{y\}$.[10] $\diamond$

Let $\mathcal{S}$ denote the set of all admissible information structures. By using the continuity of DM i's expected payoff with respect to Y_i (in the discrete metric), it is possible to show that an optimal strategy of the augmented choice problem (G, S_i) exists for every $S_i \in \mathcal{S}$, even though it may not be unique.

Proposition 1. (*Existence of optimal strategy of the augmented choice problem* (G, S_i)) The augmented choice problem (G, S_i) admits an optimal strategy, $\mathcal{P}^i_{Y|X,\epsilon,T}$, for every $S_i \in \mathcal{S}$. $\diamond$

Before concluding, we remark that the framework above nests various discrete choice models that have been analysed in the literature, as clarified by the following examples.

[9] In the presence of ties.

[10] See Appendix A for further notes on Definition 1. We explain how the product $P^i_{T|X,\epsilon,V}(t|x, e, v) P_{V|X,\epsilon}(v|x, e)$ in the inequality of Definition 1 relates to DM i's posterior. We also provide an equivalent definition of an optimal strategy of the augmented choice problem (G, S_i). Finally, we use Definition 1 to formally define DM i's endogenous consideration set.

Example 1. As a first example, we consider the Nested Logit model with one nest collecting all goods but the outside option. The payoff function, u, is

$$u(y, Z_i, \xi_i, \eta_i) \equiv \begin{cases} \beta' Z_{iy} + \lambda \log(\xi_i) + \lambda \eta_{iy} & \text{if } y \in \mathcal{Y} \setminus \{0\}, \\ \eta_{i0} & \text{if } y = 0, \end{cases} \tag{1}$$

where $0 \in \mathcal{Y}$ is the outside option, $\mathcal{Y}$ has cardinality $L+1$, $Z_i \equiv (Z_{i,1}, ..., Z_{i,L})$ is an $M \times L$ matrix of inside goods' characteristics, ξ_i and $\eta_i \equiv (\eta_{i,0}, ..., \eta_{i,L})$ represent DM i's tastes, and $\lambda \in (0,1)$. $\{\xi_i, \eta_{i,0}, ..., \eta_{i,L}\}$ are mutually independent and independent of Z_i. The densities of ξ_i and η_{iy} are parameterised as in Cardell (1997), so that $\rho_{iy} \equiv \lambda \log(\xi_i) + \lambda \eta_{iy}$ has a standard Gumbel density and the CDF of $(\rho_{i1}, ..., \rho_{iL})$ evaluated at $(s_1, ..., s_L)$ is $\exp(-(\sum_{y=1}^{L} \exp(-s_y/\lambda))^{\lambda})$.[11] Product specific dummies can enter the payoff function but we omit them for simplicity of exposition. The researcher observes the choice made by DM i and the realisation of Z_i.

In the standard Nested Logit model, every agent is assumed to know the realisation of all the payoff-relevant variables before choosing, i.e., every agent is endowed with the complete information structure. However, in some settings, the researcher may prefer to impose weaker conditions on DM i's awareness and impose that DM i observes the realisation of (Z_i, ξ_i) but might be uncertain about the realisation of η_i. For example, if $\mathcal{Y}$ collects transport modes, η_i could represent tastes for pro-environment and comfort features. Hence, following our general notation, $X_i \equiv Z_i$, $\epsilon_i \equiv \xi_i$, and $V_i \equiv \eta_i$. DM i has a prior on V_i conditional on (X_i, ϵ_i), which we assume to obey the Nested Logit parameterisation above. Further, DM i may process additional information to refine her prior by, for instance, investigating the technical characteristics of each transport mode, seeking out reviews, etc. Such additional information constitutes DM i's information structure and is left completely unrestricted when studying identification of (β, λ).

The following scenario is also possible. For each $y \in \mathcal{Y}$, let $Z_{i,y}^1$ and $Z_{i,y}^{-1}$ denote the first component and the residual $M-1$ components of the $M \times 1$ vector $Z_{i,y}$, respectively. Further, let $Z_i^1 \equiv (Z_{i,1}^1, ..., Z_{i,L}^1)$ and $Z_i^{-1} \equiv (Z_{i,1}^{-1}, ..., Z_{i,L}^{-1})$. In some settings, it may be reasonable to assume that DM i observes the realisation of $(Z_i^{-1}, \xi_i, \eta_i)$ but might be uncertain about the realisation of Z_i^1. For example, if $\mathcal{Y}$ collects transport modes, Z_i^1 could represent the associated pollution costs. Hence, following our general notation, $X_i \equiv Z_i^{-1}$, $\epsilon_i \equiv (\xi_i, \eta_i)$, and $V_i \equiv Z_i^1$. DM i has a prior on V_i conditional on (X_i, ϵ_i), which we assume to be equal to the empirical distribution of Z_i^1 conditional on Z_i^{-1}. Further, DM i may process additional information to refine her prior and form a posterior by, for instance, calculating the amount of fuel used by each transport mode, checking the road and traffic conditions of her home-work journey, etc. As above, such additional information constitutes DM i's information structure and is left completely unrestricted when studying identification of (β, λ).

Similar considerations can be made for other standard discrete choice models, e.g., Multinomial Logit/Probit model, Mixed Logit model, etc.

[11] See also Galichon (2019) regarding the random utility representation of the Nested Logit model.

Example 2. As a second example, we consider a discrete choice model of car insurance plans. Specifically, DM i faces an underlying risk of a loss (car accident) and can choose among L insurance plans. The loss event is denoted by C_i. $C_i = 1$ if the loss event occurs, and 0 otherwise. Each insurance plan $y \in \mathcal{Y}$ is characterised by a deductible, D_y, and a premium, P_{iy}. Further, DM i is endowed with some wealth (Wealth$_i$) and an $M \times 1$ vector of characteristics, Z_i, such as gender, age, insurance score, rating territories, etc. C_i is a function of Z_i and some shock η_i. For example, we can use a simple Probit model and set $C_i = \mathbb{1}\{Z_i'\beta + \eta_i \geq 0\}$, where η_i is distributed as a standard normal independently of all the other variables in the model. The payoff function, u, belongs to the CARA family, i.e., for each $y \in \mathcal{Y}$,

$$u(y, P_i, D, \text{Wealth}_i, r_i, C_i) \equiv \begin{cases} \frac{1-\exp[-\epsilon_i \times (\text{Wealth}_i - P_{i,y} - D_y)]}{\epsilon_i} & \text{if } C_i = 1, \epsilon_i \neq 0, \\ \frac{1-\exp[-\epsilon_i \times (\text{Wealth}_i - P_{i,y})]}{\epsilon_i} & \text{if } C_i = 0, \epsilon_i \neq 0, \\ W_i - P_{i,y} - D_y & \text{if } C_i = 1, \epsilon_i = 0, \\ W_i - P_{i,y} & \text{if } C_i = 0, \epsilon_i = 0, \end{cases} \tag{2}$$

where $P_i \equiv (P_{i1}, ..., P_{iL})$, $D \equiv (D_1, ..., D_L)$, and r_i is the coefficient of absolute risk aversion. The researcher observes the choice made by DM i and the realisation of $(P_i, D, \text{Wealth}_i, Z_i)$. In some cases, the researcher also observes the realisation of C_i from ex-post data on claims. Such data can be used to directly identify β.

Before choosing, DM i is aware of the realisation of $(P_i, D, \text{Wealth}_i, r_i, Z_i)$. However, DM i does not observe the realisation of (η_i, C_i) because they are realised after the insurance plan choice has been made. Hence, following our general notation, $X_i \equiv (P_i, D, \text{Wealth}_i, Z_i)$, $\epsilon_i \equiv r_i$, and $V_i \equiv (\eta_i, C_i)$. DM i has a prior on V_i conditional on (X_i, ϵ_i), which we assume to obey the Probit parameterisation above. Further, DM i may process additional information to refine her prior and form a posterior by, for instance, checking the technical features of her car, testing driving abilities of family members, verifying road and traffic conditions of daily journeys, etc. Such additional information represents DM i's information structure. Our methodology permits to study identification of the distribution of ϵ_i and β without imposing any restriction on agents' information structures.

Note that we can also allow for the case where DM i may partially observe the realisation of Z_i. This can happen, for instance, if some variables in Z_i are highly technical and require professional skills to be collected, e.g., insurance score and rating territories. Such variables should be incorporated into V_i.

Seminal papers in the literature have studied identification of agents' risk aversion and beliefs in the framework above under various assumption (Abaluck and Gruber, 2011; Barseghyan, Molinari, O' Donoghue, and Teitelbaum, 2013a; 2013b; Barseghyan, Molinari, and Teitelbaum; 2016). A typical restriction consists of parameterising (e.g., via the Poisson-Gamma credibility model) the posterior beliefs used by agents to compute expected payoffs. Here, instead, we allow agents to compute expected payoffs with any Bayes-plausible posterior.

Example 3 As a third example, we consider the rational inattention framework in Caplin and Dean (2015) and Matějka and McKay (2015). In that setting, the decision problem has two stages. In the first stage, DM i rationally chooses an information structure to update her prior. Although DM i is free to choose any information structure, attention is a scarce resource and, thus, there is a cost of processing information. As a result, more informative signal distributions are more costly. Such attentional costs are parameterised in various ways, e.g., the Shannon entropy (Sims, 2003), the posterior-separable function (Caplin, Dean, and Leahy, 2019a), etc. Formally, in the first stage DM i chooses $S_i \in \mathcal{S}$ such that

$$S_i \in \text{argmax}_{S \equiv (\mathcal{T}, \mathcal{P}_{T|X,\epsilon,V}) \in \mathcal{S}} \int_{(v,t) \in \mathcal{V} \times \mathcal{T}} \Big[\max_{y \in \mathcal{Y}} \mathbb{E}_{S,t} u(y, x_i, e_i, V_i) \Big] P_{T|X,\epsilon,V}(t|x_i, e_i, v) P_{V|X,\epsilon}(v|x_i, e_i) - C(S),$$

where $\mathbb{E}_{S,t} u(y, x_i, e_i, V_i)$ is the expected payoff from choosing $y \in \mathcal{Y}$ when DM i processes information structure S to get the posterior and receives the private signal realisation t, $C(S)$ represents the parameterised attentional costs associated with information structure S, and (x_i, e_i) are the realisations of (X_i, ϵ_i) assigned by nature to DM i. In the second stage, DM i receives the private signal realisation t_i drawn by nature according to S_i. Then, DM i chooses alternative $y \in \mathcal{Y}$ maximising $\mathbb{E}_{S_i,t_i} u(y, x_i, e_i, V_i)$. Our model nests this framework because it does not impose restrictions on attentional costs and, more generally, on how agents choose information structures.

Further, Hébert and Woodford (2018) and Morris and Strack (2019) consider continuous-time models of sequential evidence accumulation and show that the resulting choice probabilities are identical to those of a static rational inattention model with posterior-separable attentional cost functions. That is, there is an equivalence between the information that is ultimately acquired in some search models and the information acquired in a static model of rational inattention, with a particular type of attentional cost functions. Therefore, our model nests also those search frameworks.

3 Identification

In this section we discuss identification of the primitives, u, $\mathcal{P}_{V|X,\epsilon}$, and $\mathcal{P}_{\epsilon|X}$, from observing the choices made by many i.i.d. DMs facing the decision problem above. We develop a methodology that remains agnostic about information structures. We proceed under the only assumption that every DM i in the population observes *at least* a realisation of (X_i, ϵ_i) is aware of the baseline choice problem G.

Let us start by formally illustrating the assumptions on the data generating process (hereafter, DGP). In what follows, superscript 0 distinguishes the *true* value of the primitives from other possible values. The rest of the notation has been introduced in Section 2.

Assumption 1. (*DGP*) The sets $\mathcal{Y}$, $\mathcal{X}$, $\mathcal{E}$, and $\mathcal{V}$ are finite and known by the researcher. Nature repeats the following procedure for $i = 1, ..., n$, in a mutually independent manner, with n large:

1. DM i is endowed by nature with a realisation, (x_i, e_i, v_i), of (X_i, ϵ_i, V_i). x_i and e_i are drawn at random from P^0_X and $P^0_{\epsilon|X}(\cdot|x_i)$, respectively. DM i observes (x_i, e_i). DM i does not observe v_i. However, DM i has a prior on V_i conditional on $(X_i, \epsilon_i) = (x_i, e_i)$, that is $P^0_{V|X,\epsilon}(\cdot|x_i, e_i)$. The researcher observes x_i. Sometimes, the researcher may also observe some or all the components of v_i. $G^0 \equiv (\mathcal{Y}, \mathcal{X}, \mathcal{E}, \mathcal{V}, u^0, \mathcal{P}^0_{V|X,\epsilon}, \mathcal{P}^0_{\epsilon|X})$ constitutes the baseline choice problem.

2. DM i processes some information structure from the set of admissible information structures, $S^0_i \equiv (\mathcal{T}^0_i, \mathcal{P}^{i,0}_{T|X,\epsilon,V}) \in \mathcal{S}$.

3. DM i faces the augmented choice problem (G^0, S^0_i).

4. DM i chooses alternative y_i from $\mathcal{Y}$ according to the notion of optimal strategy of the augmented choice problem (G^0, S^0_i) provided in Definition 1. If needed, DM i adopts some tie-breaking rule. The researcher observes y_i.

$\diamond$

Assumption 1 summarises the model of Section 2 and draws attention to the fact that the researcher is not aware of agents' information structures and tie-breaking rules, which can be different across agents and arbitrarily correlated with the payoff-relevant variables. Among others and as mentioned before, this implies that agents compute expected payoffs with any potentially endogenous and heterogenous Bayes-plausible posteriors.

The probability mass function of (Y_i, X_i) which results from the decision problem is denoted by $P^0_{Y,X} \in \Delta(\mathcal{X} \times \mathcal{Y})$. $P^0_{Y,X}$ is nonparametrically identified by the sampling process and, hence, treated as known in the identification analysis.

In certain settings, some or all the components of v_i are observed by the researcher, together with (x_i, y_i) for $i = 1, ..., n$. For example, in discrete choice models of insurance plans, the researcher typically observes the ex-post claim experience of the agents in the sample. In those cases, $\mathcal{P}^0_{V|X,\epsilon}$ could be identified directly from such additional data under further assumptions, as highlighted in Example 2 of Section 2. In our general discussion below, for simplicity we treat v_i as unobserved by the researcher for $i = 1, ..., n$.

The sets $\mathcal{X}$, $\mathcal{E}$, and $\mathcal{V}$ are assumed finite in order to make the construction of the sharp identified set computationally tractable. Specifically, finiteness ensures that the linear programming problem in Proposition 3 has a finite number of constraints. When this is not the case, one can discretise those sets as is common in the empirical literature with partially identified parameters.

The fact that the sets $\mathcal{Y}$, $\mathcal{X}$, $\mathcal{E}$, and $\mathcal{V}$ are finite implies that the image sets of u^0, $\mathcal{P}^0_{V|X,\epsilon}$, and $\mathcal{P}^0_{\epsilon|X}$ are finite. Hence, assuming that the sets $\mathcal{Y}$, $\mathcal{X}$, $\mathcal{E}$, and $\mathcal{V}$ are finite corresponds to *parameterising* u^0, $\mathcal{P}^0_{V|X,\epsilon}$, and $\mathcal{P}^0_{\epsilon|X}$. In particular, u^0 is fully characterised by $|\mathcal{Y}|\cdot|\mathcal{X}|\cdot|\mathcal{E}|\cdot|\mathcal{V}|$ image points. Similarly, $P^0_{\epsilon|X}(\cdot|x)$ is fully characterised by $|\mathcal{E}|$ image points for each $x\in\mathcal{X}$ and $P^0_{V|X,\epsilon}(\cdot|x,e)$ is fully characterised by $|\mathcal{V}|$ image points for each $(x,e)\in\mathcal{X}\times\mathcal{E}$. We denote by $\theta^0\in\Theta\subset\mathbb{R}^K$ the vector collecting *all such image points*, with length $K\equiv|\mathcal{Y}|\cdot|\mathcal{X}|\cdot|\mathcal{E}|\cdot|\mathcal{V}|+|\mathcal{E}|\cdot|\mathcal{X}|+|\mathcal{V}|\cdot|\mathcal{E}|\cdot|\mathcal{X}|$. θ^0 is the vector of primitives that we want to identify. If the elements of the sets $\mathcal{Y}$, $\mathcal{X}$, $\mathcal{E}$, and $\mathcal{V}$ are vectors or matrices with large dimensions, then K is large too. In this case, to reduce the computational burden of the procedure, one may want to parameterise further u^0, $\mathcal{P}^0_{V|X,\epsilon}$, and $\mathcal{P}^0_{\epsilon|X}$, by assuming that u^0 is governed by the vector of parameters θ^0_1, $\mathcal{P}^0_{\epsilon|X}$ belongs to parametric family of probability distributions indexed by θ^0_2, and $\mathcal{P}^0_{V|X,\epsilon}$ belongs to parametric family of probability distributions indexed by θ^0_3. This is what we do in our simulations and empirical application. We continue the analysis without considering such additional parameterisation.

Note that Assumption 1 allows for arbitrary correlation between X_i and (ϵ_i,V_i). This is not surprising. In fact, assuming that the sets $\mathcal{X}$, $\mathcal{E}$, and $\mathcal{V}$ are finite corresponds to parameterising any correlation between X_i and (ϵ_i,V_i).

Before proceeding, let us introduce some useful notation. In what follows, we denote by $\mathcal{P}^0_{Y|X}$ the family of probability mass functions of Y_i conditional on every realisation x of X_i induced by $P^0_{Y,X}$. For each $x\in\mathcal{X}$ and $P_{Y|X}(\cdot|x)\in\Delta(\mathcal{Y})$, let us rearrange the one-to-one image set of the mapping $y\in\mathcal{Y}\mapsto P_{Y|X}(\cdot|x)\in\Delta(\mathcal{Y})$ into a $|\mathcal{Y}|\times 1$ dimensional vector. With some abuse of notation, let us still denote such a vector by $P_{Y|X}(\cdot|x)$. For each $x\in\mathcal{X}$ and $P_{\epsilon|X}(\cdot|x)\in\Delta(\mathcal{E})$, let us rearrange the one-to-one image set of the mapping $e\in\mathcal{E}\mapsto P_{\epsilon|X}(\cdot|x)\in\Delta(\mathcal{E})$ into a $|\mathcal{E}|\times 1$ dimensional vector. With some abuse of notation, let us still denote such a vector by $P_{\epsilon|X}(\cdot|x)$. Further, let us still denote the collection of these vectors across all $x\in\mathcal{X}$ by $\mathcal{P}_{\epsilon|X}$. Similarly, for each $(x,e)\in\mathcal{X}\times\mathcal{E}$ and $P_{V|X,\epsilon}(\cdot|x,e)\in\Delta(\mathcal{V})$, let us rearrange the one-to-one image set of the mapping $v\in\mathcal{V}\mapsto P_{V|X,\epsilon}(\cdot|x,e)\in\Delta(\mathcal{V})$ into a $|\mathcal{V}|\times 1$ dimensional vector. With some abuse of notation, let us still denote such a vector by $P_{V|X,\epsilon}(\cdot|x,e)$. Further, let us still denote the collection of these vectors across all $(x,e)\in\mathcal{X}\times\mathcal{E}$ by $\mathcal{P}_{V|X,\epsilon}$. Lastly, let us rearrange the one-to-one image set of the mapping $(y,x,e,v)\in\mathcal{Y}\times\mathcal{X}\times\mathcal{E}\times\mathcal{V}\mapsto u(y,x,e,v)$ into a $|\mathcal{Y}|\cdot|\mathcal{X}|\cdot|\mathcal{E}|\cdot|\mathcal{V}|\times 1$ dimensional vector. With some abuse of notation, let us still denote such a vector by u. $\theta\equiv(u,\mathcal{P}_{\epsilon|X},\mathcal{P}_{V|X,\epsilon})$ of length K represents a generic element of Θ. $S\equiv(\mathcal{T},\mathcal{P}_{T|X,\epsilon,V})$ represents a generic element of $\mathcal{S}$. Given $\theta\in\Theta$, we denote by $G^\theta\equiv(\mathcal{Y},\mathcal{X},\mathcal{E},\mathcal{V},u,\mathcal{P}_{V|X,\epsilon},\mathcal{P}_{\epsilon|X})$ the corresponding baseline choice problem.

Given the absence of restrictions on information structures and tie-breaking rules, the model is incomplete in the sense of Tamer (2003). This raises the possibility of partial identification of θ^0 and, consequently, the challenge of tractably characterising the set of θs exhausting all the implications of the model and data, i.e., the sharp identified set for θ^0.

Intuitively, the sharp identified set for θ^0 is the set of θs for which the model predicts a probability mass function of (Y_i, X_i) that matches with $P^0_{Y,X}$. More formally, for each $\theta \in \Theta$ and $S \in \mathcal{S}$, let $\mathcal{R}^{\theta,S}$ be the collection of optimal strategies of the augmented choice problem (G^θ, S). Lastly, for each $\theta \in \Theta$ and $x \in \mathcal{X}$, let $\bar{\mathcal{R}}^\theta_{Y|x}$ be the collection of probability mass functions of Y_i conditional on $X_i = x$ that are induced by the model's optimal strategies under θ, while remaining agnostic about information structures. That is,

$$\begin{aligned}\bar{\mathcal{R}}^\theta_{Y|x} \equiv \text{Conv}\Big\{ & P_{Y|X}(\cdot|x) \in \Delta(\mathcal{Y}) : \\ & P_{Y|X}(y|x) = \int_{(t,v,e)\in\mathcal{T}\times\mathcal{V}\times\mathcal{E}} P_{Y|X,\epsilon,T}(y|x,e,t) P_{T|X,\epsilon,V}(t|x,e,v) P_{V|X,\epsilon}(v|x,e) P_{\epsilon|X}(e|x) d(t,v,e)\ \forall y \in \mathcal{Y}, \\ & \mathcal{P}_{Y|X,\epsilon,T} \in \mathcal{R}^{\theta,S}, S \in \mathcal{S}\Big\},\end{aligned} \tag{3}$$

where we have used the fact that Y_i is independent of V_i conditional on (X_i, ϵ_i, T_i), because DM i's information about V_i is fully captured by T_i. Convexification (via the convex hull operator, $\text{Conv}\{\cdot\}$) allows us to include in $\bar{\mathcal{R}}^\theta_{Y|x}$ probability mass functions of Y_i conditional on $X_i = x$ that are *mixtures* across information structures and tie-breaking rules. Importantly, this implies that information structures and tie-breaking rules can vary across agents in our population. It follows that the sharp identified set for θ^0 can be defined as

$$\Theta^* \equiv \{\theta \in \Theta : P^0_{Y|X}(\cdot|x) \in \bar{\mathcal{R}}^\theta_{Y|x}\ \forall x \in \mathcal{X}\}. \tag{4}$$

Unfortunately, the definition of Θ^* in (4) seems hardly useful in practice. This is because computing $\bar{\mathcal{R}}^\theta_{Y|x}$ is intractable due to the necessity of exploring the large class $\mathcal{S}$. In what follows, we overcome such an issue by exploiting Theorem 1 in Bergemann and Morris (2016) to rewrite $P^0_{Y|X}(\cdot|x) \in \bar{\mathcal{R}}^\theta_{Y|x}$ in an equivalent but tractable way. Our discussion proceeds in four steps. First, we give the definition of one-player Bayes Correlated Equilibrium (hereafter, 1BCE) for the baseline choice problem G^θ, provided in Bergemann and Morris (2013; 2016).[12] Second, we highlight that the set of 1BCEs of the baseline choice problem G^θ is convex. Third, we introduce Theorem 1 in Bergemann and Morris (2016) which claims that the set of 1BCEs of the baseline choice problem G^θ is equivalent to the collection of optimal strategies of the augmented choice problem (G^θ, S) across every possible information structure $S \in \mathcal{S}$. Fourth, we combine the second and third steps to rewrite $P^0_{Y|X}(\cdot|x) \in \bar{\mathcal{R}}^\theta_{Y|x}$ as a simple linear programming problem.

Let us give the definition of 1BCE of the baseline choice problem G^θ.

Definition 2. (*1BCE of the baseline choice problem G^θ*) Fix $\theta \equiv (u, \mathcal{P}_{\epsilon|X}, \mathcal{P}_{V|X,\epsilon}) \in \Theta$. A

[12]Note that the notions of Bayesian Persuasion by Kamenica and Gentzkow (2011) and 1BCE coincide. Specifically, Kamenica and Gentzkow (2011) consider a framework where a sender chooses an information structure, $S \in \mathcal{S}$, to give to a receiver and then the receiver chooses an alternative. Instead of letting the sender choose an $S \in \mathcal{S}$, Bergemann and Morris (2019) finds that this is equivalent to letting the sender choose her favourite 1BCE.

family of probability mass functions of (Y_i, V_i) conditional on every realisation (x, e) of (X_i, ϵ_i),

$$\mathcal{P}_{Y,V|X,\epsilon} \equiv \{P_{Y,V|X,\epsilon}(\cdot|x,e) \in \Delta(\mathcal{Y} \times \mathcal{V}) : x \in \mathcal{X}, e \in \mathcal{E}\},$$

is a 1BCE of the baseline choice problem G^θ if:

1. It is *consistent* with the baseline choice problem G^θ, i.e., when integrating $P_{Y,V|X,\epsilon}(\cdot|x,e)$ with respect to Y_i, one obtains the prior, $P_{V|X,\epsilon}(\cdot|x,e)$, $\forall x \in \mathcal{X}$ and $\forall e \in \mathcal{E}$. That is,

$$\sum_{y \in \mathcal{Y}} P_{Y,V|X,\epsilon}(y, v|x, e) = P_{V|X,\epsilon}(v|x, e) \ \forall x \in \mathcal{X}, \forall e \in \mathcal{E}, \forall v \in \mathcal{V}.$$ [13]

2. It is *obedient*, i.e., an agent who is recommended alternative $y \in \mathcal{Y}$ by an omniscient mediator has no incentive to deviate. That is,

$$\sum_{v \in \mathcal{V}} u(y, x, e, v) P_{Y,V|X,\epsilon}(y, v|x, e) \geq \sum_{v \in \mathcal{V}} u(\tilde{y}, x, e, v) P_{Y,V|X,\epsilon}(y, v|x, e),$$
$$\forall y \in \mathcal{Y}, \forall \tilde{y} \in \mathcal{Y} \setminus \{y\}, \forall x \in \mathcal{X}, \forall e \in \mathcal{E}.$$

$\diamond$

Note that, for each $x \in \mathcal{X}$ and $e \in \mathcal{E}$, the collection of conditional probability mass functions $P_{Y,V|X,\epsilon}(\cdot|x,e)$ satisfying the consistency and obedience requirements of Definition 2 is convex. This is because the consistency and obedience requirements are linear in $P_{Y,V|X,\epsilon}(\cdot|x,e)$.

We now illustrate Theorem 1 in Bergemann and Morris (2016).[14]

Theorem 1. (*Bergemann and Morris, 2016*) Fix $\theta \equiv (u, \mathcal{P}_{\epsilon|X}, \mathcal{P}_{V|X,\epsilon}) \in \Theta$. $\mathcal{P}_{Y,V|X,\epsilon}$ is a 1BCE of the baseline choice problem G^θ if and only if there exists an information structure, $S \equiv (\mathcal{T}, \mathcal{P}_{T|X,\epsilon,V}) \in \mathcal{S}$, and an optimal strategy, $\mathcal{P}_{Y|X,\epsilon,T}$, of the augmented choice problem (G^θ, S), such that $\mathcal{P}_{Y,V|X,\epsilon}$ is induced by $\mathcal{P}_{Y|X,\epsilon,T}$.[15] $\diamond$

Note that Theorem 1 also implies that a 1BCE of the baseline choice problem G^θ exists for each $\theta \in \Theta$. Indeed, fix any information structure $S \in \mathcal{S}$. Let $\mathcal{P}_{Y|X,\epsilon,T}$ be an optimal strategy of the augmented choice problem (G^θ, S), which exists by Proposition 1. Let $\mathcal{P}_{Y,V|X,\epsilon}$ be the family of probability mass functions of (Y_i, V_i) conditional on every realisation (x, e) of (X_i, ϵ_i)

[13] Note that consistency requires that DM i applies the Bayes rule correctly to update her prior.

[14] Recall that Theorem 1 in Bergemann and Morris (2016) is valid for a general n-player game, where $n \geq 1$. It is used here for a one-player game. Note that, in a one-player game, the notion of Bayes Correlated Equilibrium does not refer to agents best responding to each other. Instead, it refers to the optimal behaviour of a single agent in a decision problem.

[15] Suppose $\mathcal{T}$ is finite. Then, by "induced" we mean

$$P_{Y,V|X,\epsilon}(y, v|x, e) = \sum_{t \in \mathcal{T}} P_{Y|X,\epsilon,T}(y|x, e, t) P_{T|X,\epsilon,V}(t|x, e, v) P_{V|X,\epsilon}(v|x, e),$$

$\forall y \in \mathcal{Y}$, $\forall v \in \mathcal{V}$, $\forall x \in \mathcal{X}$, and $\forall e \in \mathcal{E}$.

induced by $\mathcal{P}_{Y|X,\epsilon,T}$. Then, by Theorem 1, $\mathcal{P}_{Y,V|X,\epsilon}$ is a 1BCE of the baseline choice problem G^θ. Therefore, the set of 1BCE of the baseline choice problem G^θ is non-empty. Furthermore, the set of 1BCE of the baseline choice problem G^θ is typically non-singleton. Indeed, if the set of 1BCE was a singleton, then information would be essentially irrelevant, i.e., a certain alternative would be optimal regardless of any extra information that agents might process.

We now exploit Theorem 1 to represent Θ^* in an equivalent but tractable way. For each $\theta \in \Theta$, let $\mathcal{Q}^\theta$ be the collection of 1BCEs of the baseline choice problem G^θ. Moreover, for each $\theta \in \Theta$ and $x \in \mathcal{X}$, let $\bar{\mathcal{Q}}^\theta_{Y|x}$ be the collection of probability mass functions of Y_i conditional on $X_i = x$ that are induced by the 1BCEs of the baseline choice problem G^θ. That is,

$$\begin{aligned}\bar{\mathcal{Q}}^\theta_{Y|x} \equiv \Big\{ & P_{Y|X}(\cdot|x) \in \Delta(\mathcal{Y}) : \\ & P_{Y|X}(y|x) = \sum_{(e,v)\in\mathcal{E}\times\mathcal{V}} P_{Y,V|X,\epsilon}(y,v|x,e)P_{\epsilon|X}(e|x) \ \forall y \in \mathcal{Y}, \mathcal{P}_{Y,V|X,\epsilon} \in \mathcal{Q}^\theta \Big\}.\end{aligned} \tag{5}$$

Note that $\bar{\mathcal{Q}}^\theta_{Y|x}$ is convex and, hence, we do not need to take the convex hull of $\bar{\mathcal{Q}}^\theta_{Y|x}$ to allow for heterogeneity of 1BCE across agents in our population. In fact, for each $x \in \mathcal{X}$ and $e \in \mathcal{E}$, $\mathcal{Q}^\theta$ is convex because the set of 1BCEs of the baseline choice problem G^θ is convex, as highlighted above. Therefore, for each $x \in \mathcal{X}$ and $\theta \in \Theta$, $\bar{\mathcal{Q}}^\theta_{Y|x}$ is also convex.

Theorem 1 implies that $\bar{\mathcal{R}}^\theta_{Y|x} = \bar{\mathcal{Q}}^\theta_{Y|x}$ $\forall x \in \mathcal{X}$ and $\forall \theta \in \Theta$. Thus, one can rewrite Θ^* by using the notion of 1BCE, as formalised in Proposition 2.

Proposition 2. (*Characterisation of* Θ^* *through the notion of 1BCE*) Let

$$\Theta^{**} \equiv \{\theta \in \Theta : P^0_{Y|X}(\cdot|x) \in \bar{\mathcal{Q}}^\theta_{Y|x} \ \forall x \in \mathcal{X}\}.$$

Under Assumption 1, $\Theta^* = \Theta^{**}$. $\diamond$

Constructing Θ^* as characterised in Proposition 2 is computationally tractable by leveraging on the convexity of $\bar{\mathcal{Q}}^\theta_{Y|x}$ for each $x \in \mathcal{X}$ and $\theta \in \Theta$. This is formalised in Proposition 3. Before presenting Proposition 3, let us introduce some useful notation. For each $(x,e) \in \mathcal{X} \times \mathcal{E}$ and $P_{Y,V|X,\epsilon}(\cdot|x,e) \in \Delta(\mathcal{Y} \times \mathcal{V})$, let us rearrange the one-to-one image set of the mapping $(y,v) \in \mathcal{Y} \times \mathcal{V} \mapsto P_{Y,V|X,\epsilon}(\cdot|x,e) \in \Delta(\mathcal{Y} \times \mathcal{V})$ into a $|\mathcal{Y}| \cdot |\mathcal{V}| \times 1$ dimensional vector. With some abuse of notation, let us still denote such a vector by $P_{Y,V|X,\epsilon}(\cdot|x,e)$. Further, let us still denote the collection of these vectors across all $(x,e) \in \mathcal{X} \times \mathcal{E}$ by $\mathcal{P}_{Y,V|X,\epsilon}$.

Proposition 3. (*Construction of* Θ^*) Fix $\theta \equiv (u, \mathcal{P}_{\epsilon|X}, \mathcal{P}_{V|X,\epsilon}) \in \Theta$. Under Assumption 1, $\theta \in \Theta^*$ if and only if the following linear programming problem has a solution with respect to $\mathcal{P}_{Y,V|X,\epsilon}$:

$$\text{[1BCE-Consistency]:} \quad \sum_{y\in\mathcal{Y}} P_{Y,V|X,\epsilon}(y,v|x,e) = P_{V|\epsilon,X}(v|e,x) \ \forall v\in\mathcal{V}, \forall e\in\mathcal{E}, \forall x\in\mathcal{X},$$

$$\text{[1BCE-Obedience]:} \quad -\sum_{v\in\mathcal{V}} P_{Y,V|X,\epsilon}(y,v|x,e)[u(y,x,e,v)-u(y',x,e,v)] \leq 0 \ \forall y\in\mathcal{Y}, \forall y'\in\mathcal{Y}\setminus\{y\}, \forall e\in\mathcal{E}, \forall x\in\mathcal{X},$$

$$\text{[1BCE-Model predictions]:} \quad P^0_{Y|X}(y|x) = \sum_{(e,v)\in\mathcal{E}\times\mathcal{V}} P_{Y,V|X,\epsilon}(y,v|x,e)P_{\epsilon|X}(e|x) \ \forall y\in\mathcal{Y}, \forall x\in\mathcal{X}.$$

◇

Note that the linear programming problem of Proposition 3 can incorporate various classes of nonparametric assumptions into $\mathcal{P}_{V|X,\epsilon}$ and $\mathcal{P}_{\epsilon|X}$, such as independence of (ϵ_i, V_i) from X_i, symmetry of marginals around zero, identical marginals, quantile restrictions, etc. These assumptions are simply added as linear constraints.

4 Simulations

As a first example, we consider the Nested Logit model introduced in Example 1 of Section 2, when $X_i \equiv Z_i$, $\epsilon_i \equiv \xi_i$, and $V_i \equiv \eta_i$ (first scenario discussed). We start by constructing the collection of choice probabilities predicted by 1BCEs for a given value of covariates and parameters. This step serves to get a preliminary understanding about the identifying power of 1BCE. In particular, we want to exclude the possibility that 1BCE rationalises every probability distribution in the unit simplex, because this would imply that 1BCE has no identifying power. We set $L = 2$, $\beta = 0$, and $\lambda = 0.5$. Further, we discretise the densities of ϵ_i and V_i to have supports $\mathcal{E} \equiv \{0.1, 1, 2, ..., 50\}$ and $\mathcal{V} \equiv \{-2, -1, ..., 6\}^3$, respectively.[16] Hereafter, we refer to this DGP as DGP1. Let $\bar{\mathcal{R}}_Y^{\beta,\lambda,\text{comp}}$ be the collection of choice probabilities induced by the model's optimal strategies when the researcher assumes that agents are endowed with the complete information structure. Let $\bar{\mathcal{R}}_Y^{\beta,\lambda,\text{deg}}$ be the collection of choice probabilities that are induced by the model's optimal strategies when the researcher assumes that agents are endowed with the degenerate information structure. Finally, recall that $\bar{\mathcal{Q}}_Y^{\beta,\lambda}$ is the collection of choice probabilities that are induced by 1BCEs, as defined in Equation (5). Figure 1 represents $\bar{\mathcal{Q}}_Y^{\beta,\lambda}$ (black region), $\bar{\mathcal{R}}_Y^{\beta,\lambda,\text{comp}}$ (red region), and $\bar{\mathcal{R}}_Y^{\beta,\lambda,\text{deg}}$ (blue region). By Theorem 1, $\bar{\mathcal{R}}_Y^{\beta,\lambda,\text{comp}}$ and $\bar{\mathcal{R}}_Y^{\beta,\lambda,\text{comp}}$ are subsets of $\bar{\mathcal{Q}}_Y^{\beta,\lambda}$. Further, note that $\bar{\mathcal{Q}}_Y^{\beta,\lambda}$ is a strict and relatively small subset of the unit simplex, which reassures us about the identifying power of 1BCE in this context.

We now move to simulate data from (1) and construct the sharp identified set for the parameters of interest as outlined by Proposition 3. We consider a DGP slightly more complicated

[16] To discretise a density, we first discretise its support and then transform the density into a probability mass function. For example, $P_\epsilon(e) = \frac{f_\epsilon(e)}{\sum_{e\in\mathcal{E}} f_\epsilon(e)}$ $\forall e \in \mathcal{E}$, where f_ϵ is ϵ_i's original density. The support should be discretised in a way such that the resulting probability mass function preserves the shape of the original density. For a list of methods to discretise densities see Bracquemond and Gaudoin (2003), Lai (2013), and Chakraborty (2015).

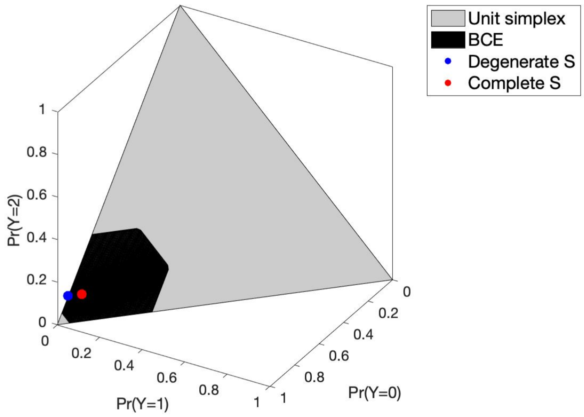

Figure 1: The figure represents $\bar{\mathcal{Q}}_Y^{\beta,\lambda}$ (black region), $\bar{\mathcal{R}}_Y^{\beta,\lambda,\text{comp}}$ (red region), and $\bar{\mathcal{R}}_Y^{\beta,\lambda,\text{deg}}$ (blue point) under DGP1.

than DGP1. In particular, to generate the data, we set $L = 3$, $M = 1$, $\beta = 1.6$, and $\lambda = 0.5$. The probability mass function of X_i, P_X, is obtained as the density of a normal random vector with mean and variance covariance matrix

$$\mu_X \equiv (0.629, 0.812, -0.746)', \; \Sigma_X \equiv \begin{pmatrix} 3.913 & 0.455 & 0.531 \\ 0.455 & 3.547 & 0.558 \\ 0.531 & 0.558 & 3.971 \end{pmatrix},$$

respectively, discretised to have support $\mathcal{X} \equiv \{-1, 0, 1\}^3$. We discretise the densities of ϵ_i and V_i to have supports $\mathcal{E} \equiv \{0.1, 1, 2, ..., 50\}$ and $\mathcal{V} \equiv \{-2, -1, ..., 6\}^3$, respectively. In the presence of ties, agents select one of the indifferent alternatives uniformly at random. Finally, the empirical choice probabilities are derived under the assumption that half of the population is endowed with the complete information structure (i.e., half of the population processes enough information to discover the exact realisation of V_i) and half of the population is endowed with the degenerate information structure (i.e., half of the population does not process additional information and has a posterior equal to the prior). Hereafter, we refer to this DGP as DGP2. The black region in Figure 2 represents the sharp identified set. The red dot in Figure 2 represents the true value of the parameters. The black region is tight and informative about the signs and magnitudes of the parameters.

Other exercises are possible. For example, the researcher can use the sharp identified set in Figure 2 to compute bounds for counterfactual choice probabilities when the researcher changes some variables in X_i and keeps (or not) information structures fixed. Details on the procedure are provided by Theorem 1 in Bergemann, Brooks, and Morris (2019). Alternatively, the researcher may be interested in constructing the sharp identified set for the second scenario discussed in Example 1 of Section 2, when $X_i \equiv Z_i^{-1}$, $\epsilon_i \equiv (\xi_i, \eta_i)$, and $V_i \equiv Z_i^1$. In turn, this set can be used to compute bounds for counterfactual choice probabilities when the researcher informs every DM i about the realisation of Z_i^1 (or about the realisation of a variable correlated

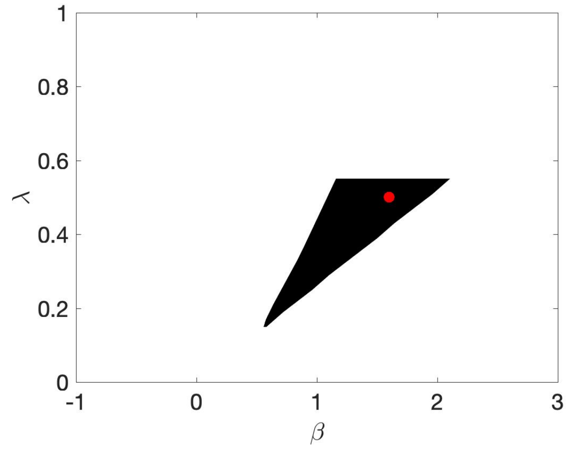

Figure 2: The figure is based on DGP2. The black region represents the sharp identified set. The red dot represents the true value of the parameters.

with Z_i^1) and keeps the payoff-relevant variables fixed. Lastly, the researcher may want to construct the sharp identified set for various minimal sets of variables observed by agents before choosing. By doing so, the researcher performs a sensitivity analysis of the identifying power of the model to changes in the minimal set of variables observed by agents before choosing.

As a second example, we consider the discrete choice model of insurance plans discussed in Example 2 of Section 2, when $X_i \equiv (P_i, D, \text{Wealth}_i, Z_i)$, $\epsilon_i \equiv r_i$, and $V_i \equiv (\eta_i, C_i)$. We start by constructing the collection of choice probabilities predicted by 1BCEs for a given value of covariates and parameters. As earlier, this step serves to get a preliminary understanding about the identifying power of 1BCE. In particular, we set $L = 3$ and $(D_1, D_2, D_3) = (100, 200, 500)$. For each insurance plan $y \in \mathcal{Y}$, $P_{i,y}$ is assumed equal to $P^{\text{base}} \times \lambda_y$, where $(\lambda_1, \lambda_2, \lambda_3) \equiv (5/6, 7/10, 3/10)$ and $P^{\text{base}} = 100$. Given that the payoff function belongs to the CARA family, payoffs can be computed without observing Wealth_i. Z_i is ignored for simplicity. ϵ_i is distributed as a Beta with parameters $\gamma_1 = 1, \gamma_2 = 10$ and support $[0, 0.02]$. Further, for computational tractability, this support is discretised into 21 equidistant points. DM i's prior on $V_i = 1$ is imposed equal to $1 - \Phi(0)$, where Φ is the normal CDF with mean 0 and variance 2. Hereafter, we refer to this DGP as DGP3. As before, let $\bar{\mathcal{R}}_Y^{\gamma_1,\gamma_2,\text{deg}}$ be the collection of choice probabilities induced by the model's optimal strategies under degenerate information structure, and $\bar{\mathcal{Q}}_Y^{\gamma_1,\gamma_2}$ is the collection of choice probabilities that are induced by 1BCEs. Figure 3 represents $\bar{\mathcal{Q}}_Y^{\gamma_1,\gamma_2}$ (black region) and $\bar{\mathcal{R}}_Y^{\gamma_1,\gamma_2,\text{deg}}$ (blue region). As earlier, $\bar{\mathcal{R}}_Y^{\gamma_1,\gamma_2,\text{deg}}$ is a subset of $\bar{\mathcal{Q}}_Y^{\gamma_1,\gamma_2}$. Further, note that $\bar{\mathcal{Q}}_Y^{\gamma_1,\gamma_2}$ is a strict and relatively small subset of the unit simplex, which reassures us about the identifying power of 1BCE in this context.

We now move to simulate data from (2) and construct the sharp identified set for the parameters of interest as outlined by Proposition 3. We consider a DGP slightly more complicated than DGP3. In particular, to generate the data, we set $L = 4$, $(D_1, D_2, D_3, D_4) = (100, 200, 500, 1000)$. For each insurance plan $y \in \mathcal{Y}$, $P_{i,y}$ is assumed equal to $P_i^{\text{base}} \times \lambda_y$, where $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) \equiv (5/6, 7/10, 3/10, 1/10)$ and P_i^{base} is uniformly distributed on $\{100, 200, 300\}$.

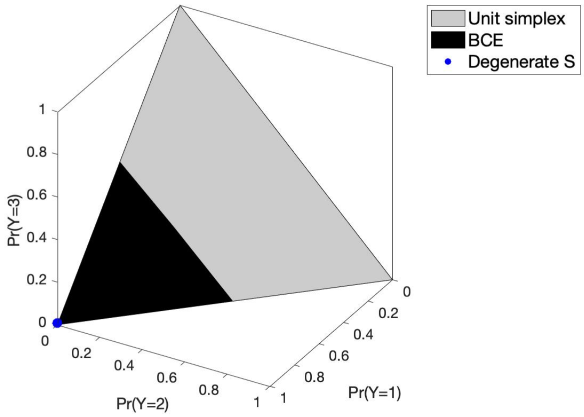

Figure 3: The figure represents $\bar{\mathcal{Q}}_Y^{\beta,\lambda}$ (black region) and $\bar{\mathcal{R}}_Y^{\beta,\lambda,\deg}$ (blue point) under DGP3.

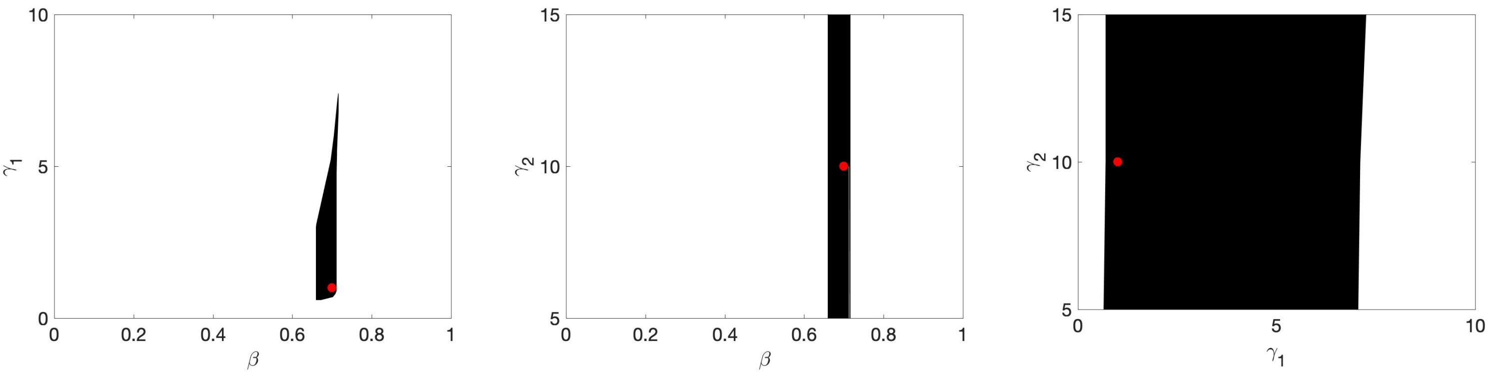

Figure 4: The figure is based on DGP4. The black regions represent the projections of the sharp identified set along each axis. The red dot represents the true value of the parameters.

Z_i is scalar and uniformly distributed on $\{-4, -3.5, -3, ..., 4\}$. ϵ_i is distributed independently of Z_i as a Beta with parameters $\gamma_1 = 1, \gamma_2 = 10$ and support $[0, 0.02]$. Further, for computational tractability, this support is discretised into 21 equidistant points. In the presence of ties, agents select one of the indifferent alternatives uniformly at random. DM i's prior on V_i being equal 1 is imposed equal to $\Phi(Z_i'\beta)$, where Φ is the normal CDF with mean 0 and variance 2 and $\beta = 0.7$. Finally, the empirical choice probabilities are derived under the assumption that the entire population processes the degenerate information structure. Hereafter, we refer to this DGP as DGP4. The black regions in Figure 4 represent the projections sharp identified set along each dimension. The red dots in Figure 4 represent the true values of the parameters. The projection for β is extremely tight. The projection for γ_1 is less tight but still bounded. The projection for γ_2 is unbounded. Note that if ex-post data on claims are available, then β can be identified directly from those. As in the previous example, other exercises are possible. For instance, the researcher can use the sharp identified set in Figure 2 to compute bounds for counterfactual choice probabilities when the researcher changes some variables in X_i and keeps (or not) information structures fixed. Alternatively, the researcher may be interested in constructing the sharp identified set when also some variables of Z_i are included in V_i because too technical to be certainly known by every DM i. In turn, this set can be used to compute bounds for counterfactual choice probabilities when the researcher informs every DM i about the realisation of the uncertain components of Z_i (or about the realisation of variables correlated with those) and keeps the payoff-relevant variables fixed. Lastly, the researcher may want to construct the sharp identified set for various sets of variables that are assumed certainly observed by agents to perform a sensitivity analysis.

5 Inference

Identification of the true parameter vector, θ_0, relies on the assumption that the true probability mass function of the observables, $P^0_{Y,X}$, is known by the researcher. However, when doing an empirical analysis, the researcher should replace $P^0_{Y,X}$ with its sample analogue resulting from having i.i.d. observations, $\{Y_i, X_i\}_{i=1}^n$, and take into account sampling variation. Given $\alpha \in (0,1)$, this section illustrates how to construct a uniformly asymptotically valid $(1-\alpha)$ confidence region, $C_{n,1-\alpha}$, for each $\theta \in \Theta^*$. In particular, we suggest to apply the generalised moment selection procedure by Andrews and Shi (2013) (hereafter, AS), as detailed in Appendix B.1 of Beresteanu, Molchanov, and Molinari (2011) (hereafter, BMM).[17] $C_{n,1-\alpha}$ is obtained by inverting, for every $\theta \in \Theta$, a test with null hypothesis $\mathrm{H}_0 : \theta_0 = \theta$. Such a test rejects H_0 if $\mathrm{TS}_n(\theta) > \hat{c}_{n,1-\alpha}(\theta)$, where $\mathrm{TS}_n(\theta)$ is a test statistic and $\hat{c}_{n,1-\alpha}(\theta)$ is a corresponding critical value. Thus, $C_{n,1-\alpha} \equiv \{\theta \in \Theta\colon\ \mathrm{TS}_n(\theta) \leq \hat{c}_{n,1-\alpha}(\theta)\}$. The remainder of the section explains how to compute $\mathrm{TS}_n(\theta)$ and $\hat{c}_{n,1-\alpha}(\theta)$ for any $\theta \in \Theta$. We anticipate that the computational advantages of using 1BCE are preserved when doing inference. Further, we remark that a similar procedure can be used to test the stability of the preference parameters across two markets, i.e., to test if changes in substitution patterns two markets are only due to changes in the distribution of covariates and/or in information structures ($\mathrm{H}_0 : \theta_0^{\text{Market 1}} - \theta_0^{\text{Market 2}} = 0_{K\times 1}$), or they can also be due to changes in the preference parameters.

First, we rewrite the linear programming of Proposition 3 as a collection of conditional moment inequalities. Let us label the elements of $\mathcal{Y}$ as $y^1, ..., y^{|\mathcal{Y}|-1}, y^{|\mathcal{Y}|}$.

Proposition 4. (*Conditional moment inequalities)* Under Assumption 1, for each $\theta \in \Theta$, $\theta \in \Theta^*$ if and only if

$$\mathbb{E}[m(Y_i, X_i; b, \theta)|X_i = x] \geq 0 \ \forall b \in \mathbb{S}^{|\mathcal{Y}|-1}, \forall x \in \mathcal{X},$$

[17] Note that the characterisation of Θ^* in Proposition 2 is equivalent to the characterisation in Theorem 2.1 of BMM. This is because the Aumann expectation of the random closed set of 1BCE alternative predictions is equal to $\bar{\mathcal{Q}}^{\theta}_{Y|x}$, for each $\theta \in \Theta$ and $x \in \mathcal{X}$.

where

$$m(Y_i, x; b, \theta) \equiv -b^T \begin{pmatrix} \mathbb{1}\{Y_i = y^1\} \\ \dots \\ \mathbb{1}\{Y_i = y^{|\mathcal{Y}|-1}\} \end{pmatrix} + \max_{P_{Y|X}(\cdot|x) \in \bar{\mathcal{Q}}^{\theta}_{Y|x}} b^T \begin{pmatrix} P_{Y|X}(y^1|x) \\ \dots \\ P_{Y|X}(y^{|\mathcal{Y}|-1}|x) \end{pmatrix}.$$

$\diamond$

Proposition 4 comes from the fact that, following BMM, one can express the condition $P^0_{Y|X}(\cdot|x) \in \bar{\mathcal{Q}}^{\theta}_{Y|x}$ as

$$-b^T P^0_{Y|X}(\cdot|x) + \sup_{P_{Y|X}(\cdot|x) \in \bar{\mathcal{Q}}^{\theta}_{Y|x}} b^T P_{Y|X}(\cdot|x) \geq 0 \ \forall b \in \mathbb{R}^{|\mathcal{Y}|}, \tag{6}$$

where the map

$$b \in \mathbb{R}^{|\mathcal{Y}|} \mapsto \sup_{P_{Y|X}(\cdot|x) \in \bar{\mathcal{Q}}^{\theta}_{Y|x}} b^T P_{Y|X}(\cdot|x) \in \mathbb{R},$$

is the support function of $\bar{\mathcal{Q}}^{\theta}_{Y|x}$. Some simple algebraic manipulations reveal that (6) is equal to the collection of conditional moment inequalities above.

Second, Lemma 2 in AS shows that conditional moment inequalities can be transformed into equivalent unconditional moment inequalities by choosing appropriate instruments, $h \in \mathcal{H}$, where $\mathcal{H}$ is a collection of instruments and h is a function of X_i. In particular,

$$\theta \in \Theta^* \Leftrightarrow \mathbb{E}[m(Y_i, X_i; b, \theta, h)] \geq 0 \ \forall b \in \mathbb{S}^{|\mathcal{Y}|-1}, \forall h \in \mathcal{H} \text{ a.s.}, \tag{7}$$

where

$$m(Y_i, X_i; b, \theta, h) \equiv m(Y_i, X_i; b, \theta) \times h(X_i).$$

Further, observe that (7) is equivalent to

$$\theta \in \Theta^* \Leftrightarrow \min\left\{0, \min_{b \in \mathbb{S}^{|\mathcal{Y}|-1}} \mathbb{E}[m(Y_i, X_i; b, \theta, h)]\right\} = 0 \ \forall h \in \mathcal{H} \text{ a.s.}$$

In light of these remarks, BMM propose as test statistic

$$\text{TS}_n(\theta) \equiv \int_{\mathcal{H}} \min\left\{0, [\min_{b \in \mathbb{S}^{|\mathcal{Y}|-1}} \sqrt{n}\bar{m}_n(b, \theta, h)]^2\right\} d\Gamma(h),$$

where Γ is a probability measure on $\mathcal{H}$ as explained in Section 3.4 of AS, and

$$\bar{m}_n(b, \theta, h) \equiv \frac{1}{n}\sum_{i=1}^{n} m(Y_i, X_i; b, \theta, h).$$

Theorem B.2 in BMM shows that, under some regularity conditions, $\text{TS}_n(\theta)$ satisfies Assumptions S1-S4 and M2 of AS. This implies that AS's procedure is applicable. Moreover, given that the set $\mathcal{X}$ is finite, the analyst can use the uniform probability measure as suggested by

Example 5 in Appendix B of AS. That is,

$$\mathrm{TS}_n(\theta) \equiv \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} \min\left\{0, [\min_{b\in\mathbb{S}^{|\mathcal{Y}|-1}} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} m(Y_i, X_i; b, \theta)\mathbb{1}\{X_i = x\}]^2\right\}. \tag{8}$$

In practice, to compute (8), the researcher should calculate, for each $x \in \mathcal{X}$,

$$\min_{b\in\mathbb{S}^{|\mathcal{Y}|-1}} \frac{1}{\sqrt{n}} \sum_{\substack{i \text{ s.t.} \\ X_i = x}} [-b^T\tilde{\mathbb{1}}_i + \max_{P_{Y|X}(\cdot|x)\in\bar{\mathcal{Q}}^{\theta}_{Y|x}} b^T\tilde{P}_{Y|X}(\cdot|x)], \tag{9}$$

where $\tilde{\mathbb{1}}_i \equiv \begin{pmatrix} \mathbb{1}\{Y_i = y^1\} \\ \vdots \\ \mathbb{1}\{Y_i = y^{|\mathcal{Y}|-1}\} \end{pmatrix}$ and $\tilde{P}_{Y|X}(\cdot|x) \equiv \begin{pmatrix} P_{Y|X}(y^1|x) \\ \vdots \\ P_{Y|X}(y^{|\mathcal{Y}|-1}|x) \end{pmatrix}$. By rearranging terms, Expression (9) becomes

$$\min_{b\in\mathbb{S}^{|\mathcal{Y}|-1}} \max_{P_{Y|X}(\cdot|x)\in\bar{\mathcal{Q}}^{\theta}_{Y|x}} b^T[-\frac{1}{\sqrt{n}} \sum_{\substack{i \text{ s.t.} \\ X_i = x}} \tilde{\mathbb{1}}_i + \frac{n_x}{\sqrt{n}}\tilde{P}_{Y|X}(\cdot|x)], \tag{10}$$

where n_x is the number of observations featuring $X_i = x$. (10) is a min-max problem which can be simplified as follows. Note that the inner constrained maximisation problem in (10) is linear in $P_{Y|X}(\cdot|x)$. Thus, it can be replaced by its dual, which consists of a linear constrained minimisation problem. Moreover, the outer constrained minimisation problem in (10) has a quadratic constraint, $b^Tb = 1$. Therefore, (10) can be rewritten as a quadratically constrained linear minimisation problem which is a tractable exercise. More details on this are in Appendix C. Once (10) is computed for each $x \in \mathcal{X}$, the analyst easily obtains $\mathrm{TS}_n(\theta)$.

To compute the critical value, we follow AS's bootstrap method consisting of the following steps. Specifically, for each $x \in \mathcal{X}$, let

$$\bar{m}_n(b, \theta, x) \equiv \frac{1}{n}\sum_{i=1}^{n} m(Y_i, X_i; b, \theta)\mathbb{1}\{X_i = x\}.$$

We draw W_n bootstrap samples using nonparametric i.i.d. bootstrap. For each $w = 1, ..., W_n$, we compute

$$TS_{n,w}(\theta) \equiv \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} \min\left\{0, [\min_{b\in\mathbb{S}^{|\mathcal{Y}|-1}} (\sqrt{n}(\bar{m}^*_{n,w}(b, \theta, x) - \bar{m}_n(b, \theta, x)) + \varphi_n(b, \theta, x))]^2\right\},$$

where $\bar{m}^*_{n,w}(b, \theta, x)$ is calculated just as $\bar{m}_n(b, \theta, x)$, but with the bootstrap sample in place of the original sample, $\varphi_n(b, \theta, h) \equiv \mathbb{1}\{\frac{1}{\kappa_n}\sqrt{n}\bar{m}_n(b, \theta, h) > 1\} \times B_n$, and $\{\kappa_n\}_{n\in\mathbb{N}}$, $\{B_n\}_{n\in\mathbb{N}}$ are sequences of constants satisfying Assumption G.1 in AS. In particular, we use $\kappa_n \equiv (0.3\log(n))^{1/2}$ and $B_n \equiv \left(\frac{0.4\log(n)}{\log(\log(n))}\right)^{1/2}$ as suggested in Section 9 of AS. Lastly, $\hat{c}_{n,1-\alpha}(\theta)$ is the $(1-\alpha)$ sample quantile of $\{TS_{n,w}(\theta)\}_{w=1}^{W_n}$.

6 Conclusions

In this paper we consider a single-agent, static, discrete choice model in which agents can face attentional limits. This implies that decision makers may be uncertain about the payoffs generated by the available alternatives. Instead of explicitly modelling the information constraints, which can be susceptible to misspecification, we study identification and inference of the preference parameters while remaining agnostic about the mechanism determining the amount of information processed by decision makers. We exploit Theorem 1 in Bergemann and Morris (2016) to provide a tractable characterisation of the sharp identified set and study inference. Simulations of discrete choice models with risk aversion and Nested Logit models highlight that our methodology can produce informative bounds for the preference parameters.

We are currently working on an empirical illustration to real data.

References

Abaluck, J., and A. Adams (2018): "What Do Consumers Consider Before They Choose? Identification from Asymmetric Demand Responses," Working Paper.

Abaluck, J., and J. Gruber (2011): "Choice Inconsistencies among the Elderly: Evidence from Plan Choice in the Medicare Part D Program," *American Economic Review*, 101(4), 1180–1210.

Abaluck, J., and G. Compiani (2019): "A Method to Estimate Discrete Choice Models that is Robust to Consumer Search," Working Paper.

Andrews, D.W.K., and X. Shi (2013): "Inference Based on Conditional Moment Inequalities," *Econometrica*, 81(2), 609–666.

Barseghyan, A., F. Molinari, T. O'Donoghue, and J.C. Teitelbaum (2013a): "The Nature of Risk Preferences: Evidence from Insurance Choices," *American Economic Review*, 103(6), 2499–2529.

Barseghyan, A., F. Molinari, T. O'Donoghue, and J.C. Teitelbaum (2013b): "Distinguishing Probability Weighting from Risk Misperceptions in Field Data," *American Economic Review: Paper & Proceedings*, 103(3), 580–585.

Barseghyan, A., F. Molinari, and J.C. Teitelbaum (2016): "Inference under Stability of Risk Preferences," *Quantitative Economics*, 7(2), 367–409.

Barseghyan, A., F. Molinari, T. O'Donoghue, and J.C. Teitelbaum (2018): "Estimating Risk Preferences in the Field," *Journal of Economic Literature*, 56(2), 501–564.

Bettinger, E.P., B.T. Long, P. Oreopoulos, L. Sanbonmatsu (2012): "The Role of Application Assistance and Information in College Decisions: Results from the H&R Block Fafsa Experiment," 127(3), 1205–1242.

Barseghyan, L., M. Coughlin, F. Molinari, and J.C. Teitelbaum (2019): "Heterogeneous Choice Sets and Preferences," arXiv:1907.02337.

Barseghyan, L., F. Molinari, and M. Thirkettle (2019): "Discrete Choice Under Risk with Limited Consideration," arXiv:1902.06629.

Beresteanu, A., I. Molchanov, and F. Molinari (2011): "Sharp Identification in Models with Convex Moment Predictions," *Econometrica*, 79(6), 1785–1821.

Bergemann, D., B. Brooks, and S. Morris (2019): "Counterfactuals with Latent Information," Cowles Foundation Discussion Paper 2162.

Bergemann, D. and S. Morris (2013): "Robust Predictions in Games With Incomplete Information," *Econometrica*, 81(4), 1251–1308.

Bergemann, D. and S. Morris (2016): "Bayes Correlated Equilibrium and the Comparison of Information Structures in Games," *Theoretical Economics*, 11(2), 487–522.

Bergemann, D. and S. Morris (2019): "Information Design: A Unified Perspective," *Journal of Economic Literature*, 57(1), 44–95.

Bracquemond, C., and O. Gaudoin (2003): "A Survey on Discrete Life Time Distributions," *International Journal of Reliability, Quality and Safety Engineering*, 10(1) 69–98.

Caplin, A., and M. Dean (2015): "Revealed Preference, Rational Inattention, and Costly Information Acquisition," *American Economic Review*, 105(7), 2183–2203.

Caplin, A., M. Dean, and J. Leahy (2019a): "Rationally Inattentive Behavior: Characterizing and Generalizing Shannon Entropy," Working Paper.

Caplin, A., M. Dean, and J. Leahy (2019b): "Rational Inattention, Optimal Consideration Sets and Stochastic Choice," *Review of Economic Studies*, 86(3), 1061–1094.

Caplin, A., D. Martin (2015): "A Testable Theory of Imperfect Perception," *The Economic Journal*, 125(582), 184–202.

Cardell, N.S. (1997): "Variance Components Structures for the Extreme-Value and Logistic Distributions with Application to Models of Heterogeneity," *Econometric Theory*, 13(2), 185–213.

Cattaneo, M., X. Ma, Y. Masatlioglu, and E. Suleymanov (2019): "A Random Attention Model," arXiv:1712.03448.

Chakraborty, S. (2015): "Generating Discrete Analogues of Continuous Probability Distributions - A Survey of Methods and Constructions," *Journal of Statistical Distributions and Applications*, 2(6).

Conlon, C.T., and J.H. Mortimer (2013): "Demand Estimation Under Incomplete Product Availability," *American Economic Journal: Microeconomics*, 5(4), 1–30.

Csaba, D. (2018): "Attentional Complements," Working Paper.

Fosgerau, M., E. Melo, A. de Palma, and M. Shum (2017): "Discrete Choice and Rational Inattention: a General Equivalence Result," arXiv:1709.09117.

Galichon, A. (2019): "On the Representation of the Nested Logit Model," arXiv:1907.08766.

Service," *American Economic Review*, 106(11), 3521–3557.

Goeree, M.S. (2008): “Limited Information and Advertising in the U.S. Personal Computer Industry,” *Econometrica*, 76(5), 1017–1074.

Hastings, J., and L. Tejeda-Ashton (2008): “Financial Literacy, Information, and Demand Elasticity: Survey and Experimental Evidence from Mexico,” NBER Working Paper 14538.

He, Y., S. Sinha, and X. Sun (2019): “Identification and Estimation in Many-to-One Two-Sided Matching without Transfers,” Working Paper.

Hébert, B., and M. Woodford (2018): “Information Costs and Sequential Information Sampling,” NBER Working Paper 25316.

Honka, E. and P. Chintagunta (2016): “Simultaneous or Sequential? Search Strategies in the US Auto Insurance Industry,” *Marketing Science*, 36(1), 21–42.

Kahneman, D. (1979): *Attention and Effort*, Prentice-Hall Englewood Cliffs, NJ.

Kamat, V. (2019): “Identification with Latent Choice Sets,” arXiv:1711.02048.

Kamenica, E., and M. Gentzkow (2011): “Basyesian Persuasion,” *American Economic Review*, 101(6), 2590–2615.

Kling, Jeffrey R., S. Mullainathan, E. Shafir, L.C. Vermeulen, M.V. Wrobel (2012): “Comparison Friction: Experimental Evidence from Medicare Drug Plans,” *Quarterly Journal of Economics*, 127(1), 199–235.

Lacetera, N., D.G. Pope, and J.R. Sydnor (2012): “Heuristic Thinking and Limited Attention in the Car Market,” *American Economic Review*, 102(5), 2206–2236.

Lai, C.D. (2013): “Issues Concerning Constructions of Discrete Lifetime Models,” *Quality Technology & Quantitative Management*, 10(2), 251–262.

Magnolfi, L. and C. Roncoroni (2017): “Estimation of Discrete Games with Weak Assumptions on Information,” Working Paper.

Manski, C. (2004): “Measuring Expectations,” *Econometrica*, 72(5), 1329–1376.

Matějka, F., and A. McKay (2015): “Rational Inattention to Discrete Choices: A New Foundation for the Multinomial Logit Model,” *American Economic Review*, 105(1), 272–298.

Mehta, N., S. Rajiv, and K. Srinivasan (2003): “Price Uncertainty and Consumer Search: A Structural Model of Consideration Set Formation,” *Marketing Science*, 22(1), 58–84.

Morris, S., and P. Strack (2019): “The Wald Problem and the Relation of Sequential Sampling and Ex-Ante Information Costs,” SSRN Working Paper.

Simon, H.A. (1955): “A Behavioral Model of Rational Choice,” *Quarterly Journal of Economics*, 69(1), 99–118.

Simon, H.A. (1959): “Theories of Decision-Making in Economics and Behavioral Science,” *American Economic Review*, 49(3), 253–83.

Sims, C.A. (1998): “Stickiness,” *Carnegie-Rochester Conference Series on Public Policy*, 49(1), 317–356.

Sims, C.A. (2003): “Implications of Rational Inattention,” *Journal of Monetary Economics* , 50(3), 665–690.

Sims, C.A. (2006): “Rational Inattention: Beyond the Linear-Quadratic Case,” *American Economic Review*, 96(2), 158–163.

Syrgkanis, V., E. Tamer, and J. Ziani (2018): “Inference on Auctions with Weak Assumptions on Information,” arXiv:1710.03830.

Tamer, E. (2003): “Incomplete Simultaneous Discrete Response Model with Multiple Equilibria,” *Review of Economic Studies*, 70(1), 147–165.

Wilson, C.J. (2008): “Consideration Sets and Political Choices: A Heterogeneous Model of Vote Choice and Sub-National Party Strength,” *Political Behavior*, 30(2), 161–183.

A Some remarks on Definition 1

We add some remarks on Definition 1. First, note that the product $P^i_{T|X,\epsilon,V}(t|x,e,v)P_{V|X,\epsilon}(v|x,e)$ in the inequality of Definition 1 is not DM i's posterior. However, by Bayes rule,

$$\int_{v\in\mathcal{V}} u(y,x,e,v)P^i_{V|X,\epsilon,T}(v|x,e,t)dv \geq \int_{v\in\mathcal{V}} u(\tilde{y},x,e,v)P^i_{V|X,\epsilon,T}(v|x,e,t)dv,$$

if and only if

$$\int_{v\in\mathcal{V}} u(y,x,e,v)P^i_{T|X,\epsilon,V}(t|x,e,v)P_{V|X,\epsilon}(v|x,e)dv \geq \int_{v\in\mathcal{V}} u(\tilde{y},x,e,v)P^i_{T|X,\epsilon,V}(t|x,e,v)P_{V|X,\epsilon}(v|x,e)dv,$$

provided that $\int_{v\in\mathcal{V}} P^i_{T|X,\epsilon,V}(t|x,e,v)P_{V|X,\epsilon}(v|x,e)dv$ is different from zero.

Second, note that we can equivalently define an optimal strategy of the augmented choice problem (G,S_i) as follows. Given $x\in\mathcal{X}$, $e\in\mathcal{E}$, and $t\in\mathcal{T}_i$, let $\mathcal{Y}^i_{x,e,t}\subseteq\mathcal{Y}$ be the set of alternatives maximising DM i's expected payoff, i.e.,

$$\mathcal{Y}^i_{x,e,t}\equiv \text{argmax}_{y\in\mathcal{Y}}\int_{v\in\mathcal{V}} u(y,x,e,v)P^i_{V|X,\epsilon,T}(v|x,e,t)dv.$$

Let $\mathcal{P}^i_{x,e,t}$ be the family of probability mass functions of Y_i conditional on $(X_i,\epsilon_i,t_i)=(x,e,t)$ that are degenerate on each of element of $\mathcal{Y}^i_{x,e,t}$. Let $\text{Conv}(\mathcal{P}^i_{x,e,t})$ be the convex hull of $\mathcal{P}^i_{x,e,t}$. Then, $\mathcal{P}^i_{Y|X,\epsilon,T}$ is an optimal strategy of the augmented choice problem (G,S_i) if $P^i_{Y|X,\epsilon,T}(\cdot|x,e,t)\in\text{Conv}(\mathcal{P}^i_{x,e,t})$ $\forall x\in\mathcal{X}$, $\forall e\in\mathcal{E}$, and $\forall t\in\mathcal{T}_i$.

Third, note that Definition 1 allows to formally defines DM i's consideration set. In fact, following Caplin, Dean, and Leahy (2019b), DM i's consideration set, $\mathcal{C}_i$, arises endogenously from $\mathcal{P}^i_{Y|X,\epsilon,T}$. In particular, $\mathcal{C}_i$ collects every alternative such that the subset of the signal's support inducing DM i to choose that alternative has positive measure. For example, when $\mathcal{T}_i$ and $\mathcal{V}$ are finite,

$$\mathcal{C}_i\equiv\{y\in\mathcal{Y}:\sum_{t\in\mathcal{T}_i}P^i_{Y|X,\epsilon,T}(y|x_i,e_i,t)\sum_{v\in\mathcal{V}}P^i_{T|X,\epsilon,V}(t|x_i,e_i,v)P_{V|X,\epsilon}(v|x_i,e_i)>0\},$$

where (x_i,e_i) are the realisations of (X_i,ϵ_i) assigned by nature to DM i.

B Proofs

Proof of Proposition 1 We proceed by construction. Take any $S_i\equiv(\mathcal{T}_i,\mathcal{P}^i_{T|X,\epsilon,V})\in\mathcal{S}$. First, note that the set $\mathcal{Y}$ is finite and, hence, compact. Second, the map $y\in\mathcal{Y}\mapsto u(y,x,e,v)\in\mathbb{R}$ is continuous using the discrete metric for each $x\in\mathcal{X}$, $e\in\mathcal{E}$, and $v\in\mathcal{V}$. Hence, the map $y\mapsto\int_{v\in\mathcal{V}}u(y,x,e,v)P^i_{T|X,\epsilon,V}(t|x,e,v)P_{V|X,\epsilon}(v|x,e)dv$ is also continuous for each $x\in\mathcal{X}$, $e\in\mathcal{E}$, and $t\in\mathcal{T}_i$. Therefore, Weierstrass theorem ensures the existence of the minimum and maximum of such a map. Given $x\in\mathcal{X}$, $e\in\mathcal{E}$, and $t\in\mathcal{T}_i$, let $y^i_{x,e,t}\in\mathcal{Y}$ be one of the maximisers. Then, an optimal strategy is $\mathcal{P}^i_{Y|X,\epsilon,T}$ such that for each $x \in \mathcal{X}$, $e \in \mathcal{E}$, and $t \in \mathcal{T}_i$,

$$P^i_{Y|X,\epsilon,T}(y^i_{x,e,t}|x,e,t) = 1 \text{ and } P^i_{Y|X,\epsilon,T}(\tilde{y}|x,e,t) = 0 \ \forall \tilde{y} \in \mathcal{Y} \setminus \{y^i_{x,e,t}\}.$$

Proof of Proposition 2 Take any $\theta \in \Theta$ and $x \in \mathcal{X}$. We show that if $P_{Y|X}(\cdot|x) \in \bar{\mathcal{Q}}^\theta_{Y|x}$, then $P_{Y|X}(\cdot|x) \in \bar{\mathcal{R}}^\theta_{Y|x}$. If $P_{Y|X}(\cdot|x) \in \bar{\mathcal{Q}}^\theta_{Y|x}$, then, by definition of $\bar{\mathcal{Q}}^\theta_{Y|x}$, there exists $\mathcal{P}_{Y,V|X,\epsilon} \in \mathcal{Q}^\theta$ inducing $P_{Y|X}(\cdot|x)$. By Theorem 1, it follows that there exists $S \in \mathcal{S}$ and $\mathcal{P}_{Y|X,\epsilon,T} \in \mathcal{R}^{\theta,S}$ such that $\mathcal{P}_{Y|X,\epsilon,T}$ induces $\mathcal{P}_{Y,V|X,\epsilon}$. Thus, $\mathcal{P}_{Y|X,\epsilon,T}$ induces $P_{Y|X}(\cdot|x)$ by the transitive property. Therefore, by definition of $\bar{\mathcal{R}}^\theta_{Y|x}$, $P_{Y|X}(\cdot|x) \in \bar{\mathcal{R}}^\theta_{Y|x}$.

Conversely, we show that $P_{Y|X}(\cdot|x) \in \bar{\mathcal{R}}^\theta_{Y|x}$, then $P_{Y|X}(\cdot|x) \in \bar{\mathcal{Q}}^\theta_{Y|x}$. First, let $\tilde{\mathcal{R}}^\theta_{Y|x} \subseteq \bar{\mathcal{R}}^\theta_{Y|x}$ be the non-convexified collection of probability mass functions of Y_i conditional $X_i = x$ that are induced by the model's optimal strategies under θ, while remaining agnostic about information structures. That is,

$$\begin{aligned}\tilde{\mathcal{R}}^\theta_{Y|x} \equiv \Big\{&P_{Y|X}(\cdot|x) \in \Delta(\mathcal{Y}) : \\ &P_{Y|X}(y|x) = \int_{\mathcal{T}\times\mathcal{V}\times\mathcal{E}} P_{Y|X,\epsilon,T}(y|x,e,t)P_{T|X,\epsilon,V}(t|x,e,v)P^\theta_{V|X,\epsilon}(v|x,e)P^\theta_{\epsilon|X}(e|x)d(t,v,e) \ \forall y \in \mathcal{Y}, \\ &\mathcal{P}_{Y|X,\epsilon,T} \in \mathcal{R}^{\theta,S}, S \in \mathcal{S}\Big\}.\end{aligned}$$

Take $P_{Y|X}(\cdot|x) \in \tilde{\mathcal{R}}^\theta_{Y|x}$. Then, by definition of $\tilde{\mathcal{R}}^\theta_{Y|x}$, there exists $S \in \mathcal{S}$ and $\mathcal{P}_{Y|X,\epsilon,T} \in \mathcal{R}^{\theta,S}$ such that $\mathcal{P}_{Y|X,\epsilon,T}$ induces $P_{Y|X}(\cdot|x)$. By Theorem 1, it follows that there exists $\mathcal{P}_{Y,V|X,\epsilon} \in \mathcal{Q}^\theta$ inducing $\mathcal{P}_{Y|X,\epsilon,T}$. Thus, $\mathcal{P}_{Y,V|X,\epsilon}$ induces $P_{Y|X}(\cdot|x)$ by the transitive property. Hence, by definition of $\bar{\mathcal{Q}}^\theta_{Y|x}$, $P_{Y|X}(\cdot|x) \in \bar{\mathcal{Q}}^\theta_{Y|x}$. Now, take any K elements from $\tilde{\mathcal{R}}^\theta_{Y|x}$, for any K. Denote such elements by $P^1_{Y|X}(\cdot|x) \in \tilde{\mathcal{R}}^\theta_{Y|x}, ..., P^K_{Y|X}(\cdot|x) \in \tilde{\mathcal{R}}^\theta_{Y|x}$. Given the arguments above, it holds that $P^1_{Y|X}(\cdot|x) \in \bar{\mathcal{Q}}^\theta_{Y|X}(\cdot|x), ..., P^K_{Y|X}(\cdot|x) \in \bar{\mathcal{Q}}^\theta_{Y|x}$. Moreover, any convex combination of $P^1_{Y|X}(\cdot|x), ..., P^K_{Y|X}(\cdot|x)$ belongs to $\bar{\mathcal{Q}}^\theta_{Y|x}$ because $\bar{\mathcal{Q}}^\theta_{Y|x}$ is convex. Therefore, every $P_{Y|X}(\cdot|x) \in \bar{\mathcal{R}}^\theta_{Y|x}$ is also contained in $\bar{\mathcal{Q}}^\theta_{Y|x}$.

We can conclude that $\bar{\mathcal{R}}^\theta_{Y|x} = \bar{\mathcal{Q}}^\theta_{Y|x}$ $\forall \theta \in \Theta$ and $\forall x \in \mathcal{X}$. This implies $\Theta^* = \Theta^{**}$.

Proof of Proposition 3 Proposition 3 is obtained by combining Proposition 2 with Definition 2 to write explicitly $\bar{\mathcal{Q}}^\theta_{Y|x}$ for each $x \in \mathcal{X}$ and $\theta \in \Theta$.

Proof of Proposition 4 Fix any $\theta \in \Theta$ and $x \in \mathcal{X}$. Observe that

$$P^0_{Y|X}(\cdot|x) \in \bar{\mathcal{Q}}^\theta_{Y|x} \ \Leftrightarrow \ -b^T P^0_{Y|X}(\cdot|x) + \sup_{P_{Y|X}(\cdot|x)\in\bar{\mathcal{Q}}^\theta_{Y|x}} b^T P_{Y|X}(\cdot|x) \geq 0 \ \forall b \in \mathbb{R}^{|\mathcal{Y}|}. \tag{B.1}$$

By the positive homogeneity of the support function, $\forall b \in \mathbb{R}^{|\mathcal{Y}|}$,

$$-b^T P^0_{Y|X}(\cdot|x) + \sup_{P_{Y|X}(\cdot|x)\in\bar{\mathcal{Q}}^\theta_{Y|x}} b^T P_{Y|X}(\cdot|x) \geq 0 \;\Leftrightarrow\; -\frac{b^T}{||b||} P^0_{Y|X}(\cdot|x) + \sup_{P_{Y|X}(\cdot|x)\in\bar{\mathcal{Q}}^\theta_{Y|x}} \frac{b^T}{||b||} P_{Y|X}(\cdot|x) \geq 0. \tag{B.2}$$

By (B.2), (B.1) is equivalent to

$$P^0_{Y|X}(\cdot|x) \in \bar{\mathcal{Q}}^\theta_{Y|x} \;\Leftrightarrow\; -b^T P^0_{Y|X}(\cdot|x) + \sup_{P_{Y|X}(\cdot|x)\in\bar{\mathcal{Q}}^\theta_{Y|x}} b^T P_{Y|X}(\cdot|x) \geq 0 \;\forall b \in \mathbb{S}^{|\mathcal{Y}|}. \tag{B.3}$$

Moreover, given that $\bar{\mathcal{Q}}^\theta_{Y|x}$ is closed and bounded, (B.3) is equivalent to

$$P^0_{Y|X}(\cdot|x) \in \bar{\mathcal{Q}}^\theta_{Y|x} \;\Leftrightarrow\; -b^T P^0_{Y|X}(\cdot|x) + \max_{P_{Y|X}(\cdot|x)\in\bar{\mathcal{Q}}^\theta_{Y|x}} b^T P_{Y|X}(\cdot|x) \geq 0 \;\forall b \in \mathbb{S}^{|\mathcal{Y}|}. \tag{B.4}$$

Lastly, given that $\bar{\mathcal{Q}}^\theta_{Y|x}$ is a subset of the $(|\mathcal{Y}|-1)$-dimensional simplex, (B.4) is equivalent to

$$P^0_{Y|X}(\cdot|x) \in \bar{\mathcal{Q}}^\theta_{Y|x} \;\Leftrightarrow\; -b^T \begin{pmatrix} P^0_{Y|X}(y^1|x) \\ \vdots \\ P^0_{Y|X}(y^{|\mathcal{Y}|-1}|x) \end{pmatrix} + \max_{P_{Y|X}(\cdot|x)\in\bar{\mathcal{Q}}^\theta_{Y|x}} b^T \begin{pmatrix} P_{Y|X}(y^1|x) \\ \vdots \\ P_{Y|X}(y^{|\mathcal{Y}|-1}|x) \end{pmatrix} \geq 0 \;\forall b \in \mathbb{S}^{|\mathcal{Y}|-1}. \tag{B.5}$$

Therefore, by combining Proposition 2 with (B.5), we get that

$$\theta \in \Theta^* \Leftrightarrow -b^T \begin{pmatrix} P^0_{Y|X}(y^1|x) \\ \vdots \\ P^0_{Y|X}(y^{|\mathcal{Y}|-1}|x) \end{pmatrix} + \max_{P_{Y|X}(\cdot|x)\in\bar{\mathcal{Q}}^\theta_{Y|x}} b^T \begin{pmatrix} P_{Y|X}(y^1|x) \\ \vdots \\ P_{Y|X}(y^{|\mathcal{Y}|-1}|x) \end{pmatrix} \geq 0 \;\forall b \in \mathbb{S}^{|\mathcal{Y}|-1}, \tag{B.6}$$

which is equivalent to

$$\theta \in \Theta^* \Leftrightarrow \mathbb{E}[m(Y_i, X_i; b, \theta | X_i = x)] \geq 0 \;\forall b \in \mathbb{S}^{|\mathcal{Y}|-1},$$

as claimed in Proposition 4.

C Inference: some computational simplifications

Fix any $\theta \in \Theta$ and $x \in \mathcal{X}$. Recall Expression (10), which we report here

$$\min_{b\in\mathbb{S}^{|\mathcal{Y}|-1}} \max_{P_{Y|X}(\cdot|x)\in\bar{\mathcal{Q}}^\theta_{Y|x}} b^T[-\frac{1}{\sqrt{n}} \sum_{\substack{i \text{ s.t.} \\ x}} \tilde{\mathbb{1}}_i + \frac{n_x}{\sqrt{n}} \tilde{P}_{Y|X}(\cdot|x)]. \tag{C.1}$$

By using Definition 2 to write explicitly the feasible set $\bar{\mathcal{Q}}^{\theta}_{Y|x}$, (C.1) is equivalent to

$$
\begin{aligned}
\min_{b\in\mathbb{R}^{|\mathcal{Y}|-1}} \quad & \max_{\substack{P_{Y|X}(\cdot|x)\,\in\,\mathbb{R}_+^{|\mathcal{Y}|} \\ P_{Y,V|X,\epsilon}(\cdot|x,e)\,\in\,\mathbb{R}_+^{|\mathcal{Y}|\cdot|\mathcal{V}|},\forall e\in\mathcal{E}}} \quad b^T[-\frac{1}{\sqrt{n}}\sum_{\substack{i \text{ s.t.} \\ x}} \tilde{\mathbb{1}}_i + \frac{n_x}{\sqrt{n}}\tilde{P}_{Y|X}(\cdot|x)],\\
\text{s.t.} \quad & [b\in\mathbb{S}^{|\mathcal{Y}|-1}]: \quad b^T b = 1,\\
& [\text{1BCE-Consistency}]: \quad \sum_{y\in\mathcal{Y}} P_{Y,V|X,\epsilon}(y,v|x,e) = P_{V|\epsilon,X}(v|e,x)\ \forall v\in\mathcal{V},\forall e\in\mathcal{E},\forall x\in\mathcal{X},\\
& [\text{1BCE-Obedience}]: \quad -\sum_{v\in\mathcal{V}} P_{Y,V|X,\epsilon}(y,v|x,e)[u(y,x,e,v)-u(y',x,e,v)]\le 0\ \forall y\in\mathcal{Y},\forall y'\in\mathcal{Y}\setminus\{y\},\forall e\in\mathcal{E},\forall x\in\mathcal{X},\\
& [\text{1BCE-model predictions}]: \quad P_{Y|X}(y|x) = \sum_{(e,v)\in\mathcal{E}\times\mathcal{V}} P_{Y,V|X,\epsilon}(y,v|x,e)P_{\epsilon|X}(e|x)\ \forall y\in\mathcal{Y},\forall x\in\mathcal{X}.
\end{aligned}
\tag{C.2}
$$

We simplify (C.2) by introducing new variables. Let

$$
\mathbb{1}_i \equiv \begin{pmatrix} \mathbb{1}\{Y_i = y^1\} \\ \vdots \\ \mathbb{1}\{Y_i = y^{|\mathcal{Y}|}\} \end{pmatrix},
$$

and

$$
Z_1 \equiv -\frac{1}{\sqrt{n}}\sum_{\substack{i \text{ s.t.} \\ X_i = x}} \mathbb{1}_i + \frac{n_x}{\sqrt{n}} P_{Y|X}(\cdot|x).
$$

Note that Z_1 is a $|\mathcal{Y}|\times 1$ vector. Further, let Z_2 be the $(|\mathcal{Y}|\cdot|\mathcal{V}|\cdot|\mathcal{E}|)\times 1$ vector collecting $P_{Y,V|X,\epsilon}(\cdot|x,e)$ across every $e\in\mathcal{E}$. Lastly, let Z be the $(|\mathcal{Y}|+|\mathcal{Y}|\cdot|\mathcal{V}|\cdot|\mathcal{E}|)\times 1$ vector collecting Z_1 and Z_2. (C.2) can be rewritten as

$$
\begin{aligned}
\min_{b\in\mathbb{R}^{|\mathcal{Y}|-1}} \quad & \max_{\substack{Z_1\,\in\,\mathbb{R}^{|\mathcal{Y}|} \\ Z_2\,\in\,\mathbb{R}_+^{|\mathcal{Y}|\cdot|\mathcal{V}|\cdot|\mathcal{E}|}}} \begin{bmatrix} b^T & 0 & 0^T_{|\mathcal{Y}|\cdot|\mathcal{V}|\cdot|\mathcal{E}|}\end{bmatrix} Z,\\
\text{s.t.} \quad & b^T b = 1,\\
& A_{\text{eq}}\ Z = B_{\text{eq}},\\
& A_{\text{ineq}}\ Z \le 0_{d_{\text{ineq}}},
\end{aligned}
\tag{C.3}
$$

where A_{eq} is the matrix of coefficients multiplying Z in the equality constraints of (C.2) with d_{eq} rows, B_{eq} is the vector of constants appearing in the equality constraints of (C.2), and A_{ineq} is the matrix of coefficients multiplying Z in the inequality constraints of (C.2) with d_{ineq} rows.

Further, the inner constrained maximisation problem in (C.3) is linear. Hence, by strong duality, can be replaced with its dual. This allows us to solve one unique minimisation problem.

Precisely, the solution of (C.3) is equivalent to the solution of

$$
\begin{aligned}
\min_{\substack{b \in \mathbb{R}^{|\mathcal{Y}|-1} \\ \lambda_{\text{eq}} \in \mathbb{R}^{d_{\text{eq}}} \\ \lambda_{\text{ineq}} \in \mathbb{R}_+^{d_{\text{ineq}}}}} \quad & \begin{bmatrix} B_{\text{eq}}^T & 0_{d_{\text{ineq}}}^T \end{bmatrix} \lambda, \\
\text{s.t.} \quad & b^T b = 1, \\
& [A^T]_{1:|\mathcal{Y}|} \lambda = \begin{pmatrix} b \\ 0 \end{pmatrix}, \\
& [A^T]_{|\mathcal{Y}|+1:|\mathcal{Y}|+|\mathcal{Y}|\cdot|\mathcal{V}|\cdot|\mathcal{E}|} \lambda \geq 0_{|\mathcal{Y}|\cdot|\mathcal{V}|\cdot|\mathcal{E}|},
\end{aligned}
\tag{C.4}
$$

where λ is the $(d_{\text{eq}} + d_{\text{ineq}}) \times 1$ vector collecting λ_{eq} and λ_{ineq}, A is the $(d_{\text{eq}} + d_{\text{ineq}}) \times (|\mathcal{Y}| + |\mathcal{Y}| \cdot |\mathcal{V}| \cdot |\mathcal{E}|)$ matrix obtained by stacking one on top of the other the matrices A_{eq} and A_{ineq}, and $[A]_{i:j}$ denotes the sub-matrix of A containing the rows $i, i+1, ..., j$ of A.

Note that (C.4) is a quadratically constrained linear minimisation problem. In particular, the first constraint in (C.4) is quadratic. The objective function and the remaining constraints in (C.4) are linear.

Close derivations are discussed in Magnolfi and Roncoroni (2017) for an entry game setting.