Identifying Heterogeneous Decision Rules From Choices When Menus Are Unobserved

Larry G. Epstein[a]
[a]Corresponding author. larry.epstein@mcgill.ca

Kaushil Patel[b]
[b]kaushil.patel@mail.mcgill.ca
[a,b]Department of Economics, McGill University
855 Sherbrooke St W., Montreal, Quebec H3A 0C4

June 1, 2025

Abstract

Consider aggregate choice data from a population with heterogeneity in both preferences (or more general decision rules) and in menus, and where the analyst has limited information about how menus are distributed across the population. We determine what can be inferred from aggregate data about the distribution of preferences by identifying the set of all distributions that are consistent with the preceding. Our main theorem strengthens and generalizes existing results on such identification and provides an alternative analytical approach (using convex capacities) to study the problem.

Keywords: discrete choice, partial identification, unobserved heterogeneity, convex capacities, core

1 Introduction

1.1 Motivation and outline

Consider the problem of explaining the distribution of choices in a heterogeneous population. Denote by λ the probability distribution of chosen alternatives, the data. A common approach is to posit heterogeneity in decision rules (or underlying preferences) and possibly also in the menus from which alternatives are chosen. A decision rule d specifies the alternative $d(A)$ chosen from each menu A; the set of all decision rules is $\mathcal{D}$. An individual with decision rule d faces menu A with probability $\pi_d(A)$. Decision rules are distributed according to a probability measure Q that is to be inferred from the data, while the collection of probability measures $\{\pi_d\}_{d\in\mathcal{D}}$ is known to the analyst (possibly up to unknown parameters).[1] Accordingly, she seeks Q satisfying, *for the given* $\{\pi_d\}_{d\in\mathcal{D}}$,

$$\lambda(a) = \sum_d \sum_A Q(d)\, \pi_d(A)\ \mathbf{1}_{d(A)=a}, \tag{1.1}$$

for all alternatives a. Then empirical frequencies are rationalized by the heterogeneity in decision rules described by Q. Of particular interest is the set of all rationalizing Qs (the sharp identified region).

The above model is general in that it covers the bulk of the discrete choice literature where various special cases are adopted;[2] for example, the traditional assumption (McFadden 1974) that the menu corresponding to each choice is observed corresponds to the special case where, for each d, $\pi_d(A_d) = 1$ for some A_d. However, data about menus that would support knowledge of the conditional probabilities π_d are often unavailable (see Manski (1977) and the overviews and many references in Barseghyan et al (2021, pp. 2016-2017, 2041-2043) and Azrieli and Rehbeck (2023)). Notably, decision models based on consideration sets (Abaluck and Adams-Prassl 2021, Cattaneo et al 2020, Manzini and Mariotti 2014, Masatlioglu et al 2012) or rational inattention (Caplin et al 2019) view choices as made from subjective menus, thus arguing against their observability.[3] One is led to the concern

[1] Filiz-Ozbay and Masatlioglu (2023) call this a random-choice model (RCM), defined by a probability distribution over a collection of choice functions (potentially irrational). They axiomatize a specific class of RCMs under the assumption of rich stochastic choice data.

[2] We are ignoring covariates that often appear in this literature, and that could be added below, because they are not germane to our contribution. We adopt a streamlined formulation in order to maximize transparency of the theoretical point of this paper.

[3] To be clear, we use "menu" to refer to the set from which an alternative is chosen

that conclusions about the identified set of measures Q that are based on (1.1) sometimes rely on ad hoc assumptions about menus.

An objective in this paper is to robustify the above model by incorporating the analyst's imperfect knowledge about menus. One alternative to the perfect information assumption is complete ignorance about menus - "anything goes" for specifications of π_ds. However, in general, one would expect there to be partial information about the menu process which, if exploited, would permit sharper identification. Therefore, we admit a range of assumptions about the analyst's information that are intermediate between complete ignorance and perfect knowledge. In all cases, we show (Theorem 2.1) that the implied identified set of distributions consists of all measures Q satisfying a finite set of linear inequalities and hence forms a polytope (a convex set with finitely many extreme points); in particular, it is computationally tractable.

Our formulation uses convex (or supermodular) capacities and their cores. (The appendix collects the few basic definitions and facts regarding capacities that are used below; a very accessible and comprehensive reference is Grabisch (2016).) Capacities are set functions that generalize probability measures in order to permit a role in the representation of beliefs for limited information and the resulting limited confidence in any single probability measure - in other terms, uncertainty about probabilities. They arise in decision theory, notably in Schmeidler's (1989) Choquet expected utility theory, where convexity of the capacity is identified with aversion to such uncertainty and where convexity characterizes the Choquet models that conform also to multiple-priors utility (Gilboa and Schmeidler 1989).[4] For our purposes, the key technical feature of convex capacities is that "the core of a mixture of capacities equals the mixture of their cores" (see (A.4) for a formal statement). Given our formulation, this property leads to a short transparent (indeed elementary) proof of our theorem that applies to and unifies all of our specifications. We view this simplicity and the associated epistemic perspective as a strength and a contribution.

The scope of our results merits emphasis. Thus far we have interpreted

by maximizing preference or by applying another decision rule. Consequently, it may be a strict subset of the objective feasible set, (for example, a consideration set), that is determined by the individual's cognitive deliberation process and is unobservable to the analyst.

[4]Convex capacities, or equivalently their conjugates, known as 2-alternating, are important also in statistical theory (in proving an extension of the Neyman-Pearson Lemma (Huber and Strassen 1973) and in supporting a version of Bayes' theorem for capacities (Wasserman and Kadane 1990). They appear also in cooperative game theory as characteristic functions. However, the epistemic interpretation is a better fit here.

the paper as addressing heterogeneity in choice assuming heterogeneity in decision rules and the unobservability of menus. However, with suitable reinterpretation of the symbols in the formal model, Theorem 2.1 *applies also to other contexts where one seeks the identification of a heterogeneous characteristic of prime interest that is robust to other unobservables.* Two such settings are (see section 3): (i) Identification of the distribution of threshold levels in a population of satisficing decision makers given their choices but where individuals differ in the order in which they consider alternatives. (ii) Identification of the distribution of effort in a population of workers who share common observable characteristics (e.g. education and experience) and who work independently, given the empirical frequency distribution of outputs but where other factors that may influence output are poorly understood.

1.2 Related literature

First we relate our contribution to some recent papers in discrete choice (and related econometrics) that also weaken a priori assumptions about menus. Barseghyan et al (2021) study identification in a random utility model where the distribution of menus in the population is unknown. Two relatively minor differences from our model are that: they assume preference maximization (particularly, Sen's α condition) rather than general decision rules, and they assume that all menus of size at least κ, ($\kappa \geq 2$), a parameter specified by the analyst, are conceivable for any individual conditional on her preference order, while we allow the set of conceivable menus to be arbitrary. (An example in the Supplementary Appendix illustrates that admitting some singleton menus can affect identification.) More importantly, they deal *only with the case of complete ignorance* of the menu process, for which their characterization of the sharp identified set corresponds (apart from their inclusion of covariates) to our complete-ignorance result in Theorem 2.1.[5] This difference from the present paper is reflected in a difference in proofs. Barseghyan et al (2021) applies the theory of random sets. This approach is limiting because, as is well known, each random set can be identified with a belief function which is a very special kind of convex capacity that precludes many of the richer information structures (those short of complete ignorance) that are accommodated in our theorem. Lu (2022) assumes that all conceivable menus are bounded above and below in the sense of set inclusion, and that the bounding menus are *known*. He uses the latter, and the assumption that decision rules satisfy Sen's α, to describe a superset of the identified region. In contrast, our conditions on Q are both necessary and sufficient for

[5]Minor differences are described following the statement of our theorem in section 2.3.

Q to rationalize the data, thus yielding the sharp identified region. Azrieli and Rehbeck (2023) also study what can be learned from aggregate choice frequencies, but with several differences from the present paper. A major difference is that they assume that the marginal empirical distributions of both menus and choices are known (constitute the data). In their study of random utility models, they assume that menus are homogeneous across decision makers, (that is, the distribution of menus does not depend on the decision rule), while we allow for correlation between menus and decision rules. Where menus are based on consideration one would expect them to depend on preference (or decision rule), as in the applied papers by Goeree (2008), and Abaluck and Adams-Prassl (2021). An example in the Supplementary Appendix shows that menu-homogeneity can strictly shrink the sharp identification region. Regarding proof arguments, they also highlight their use of "known properties of the core," though these do not not include the key property that we exploit here, and they borrow more from cooperative game theory than from decision theory and thus do not emphasize epistemics in their interpretations. Further, their proofs (specifically for their Proposition 9) use not only core properties but also network flow arguments (based on a version of Hall's marriage theorem), while we use only the single mixture property of the core noted above.

Dardanoni et al (2020) also explore what can be inferred from aggregate choice data, though their focus is on cognitive heterogeneity rather than on preference (or decision rule) heterogeneity. Individuals differ in cognitive "type" and, given an objective feasible set, they arrive at different consideration sets (menus in our terminology); further, they do so in a way that conforms to specific functional forms - the "consideration capacity model" (which limits the cardinality of the consideration set) or the "consideration probability model" (Manzini and Mariotti 2014). In the section most closely related to our paper, where preferences are unobservable and heterogeneous, they assume that choices are observed from multiple "occasions" across which both the feasible set and cognitive heterogeneity are stable. With this rich dataset and functional form restrictions they prove point identification of the distribution of cognitive types in the consideration capacity model. Roughly speaking, from the perspective of our formal framework, they severely restrict the distribution of decision rules and aim at identification of the menu formation process (π_d), which reflects the distribution of cognitive type. Unsurprisingly, their proof arguments are much different than ours.

Doval and Eilat (2023) study the setting where the analyst knows the marginal over an agent's actions and the prior over states of the world, but does not know the distribution of actions given realizations of the states of the world. They ask when two such marginals (over actions and states,

similar to the dataset in Azrieli and Rehbeck 2023) can be rationalized (in the sense of a Bayes correlated equilibrium) as the outcome of the agent learning something about the state before taking an action. Their characterization result is two systems of linear inequalities that are necessary and sufficient for the dataset to be consistent with a Bayes correlated equilibrium. One of these can be established using our "mixture of cores" property. Their proof relies partly on network flow arguments.

Galichon and Henry (2011) are, to the best of our knowledge, the first to demonstrate the usefulness of convex capacities for characterizing partially identified sets. Their context, which differs from ours, is the identification of structural parameters in models with normal form games having multiple mixed strategy equilibria and where little is understood about selection. Another difference is that they do not use the "mixture of cores" property that is central to this paper.

2 Robust identification

2.1 Preliminaries

The (finite) universal set of alternatives is X, and the set of probability distributions or measures on X is denoted $\Delta(X)$. Each individual in a finite population faces a menu, a subset of X, from which she chooses one alternative. The collection of all "relevant" menus is denoted $\mathcal{A}$, with generic element A. The collection $\mathcal{A}$ is a primitive, determined by the analyst. Another primitive is a finite set $\mathcal{D}$ of decision rules, where, for each d in $\mathcal{D}$, $d(A)$ denotes the alternative that d chooses from the menu A in $\mathcal{A}$. We do not impose any requirements on decision rules, for example, they need not be derived from preference maximization. The data to be explained are represented by $\lambda \in \Delta(X)$, the empirical frequency distribution of chosen alternatives across the population.

The analyst's view of the menu formation process determines what constitutes an "explanation." We assume that she is certain that only menus in $\mathcal{A}$ are relevant, but otherwise she has limited understanding of how menus are determined; in particular, she cannot be confident in any single conditional probability distribution over menus $\pi_d \in \Delta(\mathcal{A})$, which suggests modeling via a *set* of conditional distributions. *These sets are not "data", but rather are subjective, chosen by the analyst, in a way that captures her limited confidence and desired robustness* much as sets of priors are interpreted in the maxmin model of decision-making (Gilboa and Schmeidler (1989)). We proceed in this way, though with a slight twist as explained next.

Let

$$C_d = \{d(A) : A \in \mathcal{A}\}. \tag{2.1}$$

Thus C_d denotes the set of all alternatives that can be chosen by d for some menu. For a given d, the analyst can be sure that an element of C_d will be selected, but since the choice depends on the menu, her limited knowledge of menus affects her view of which choice is associated with d. For any given distribution over menus $\pi_d \in \Delta(\mathcal{A})$, induced beliefs over alternatives are given by $\rho_d \in \Delta(C_d)$, where

$$\rho_d(a) = \pi_d(\{A \in \mathcal{A} : d(A) = a\}), \text{ for every } a \in X. \tag{2.2}$$

Using ρ_d, (1.1) implies that

$$\lambda(a) = \sum_d Q(d)\,\rho_d(a), \quad \text{for every } a \in X, \tag{2.3}$$

where menus have been eliminated and distributions over alternatives are described by both the empirical measure λ and by the "explanatory" measures $\{\rho_d\}$. To proceed, we adopt as the benchmark notion of an explanation of λ that "(2.3) is satisfied by $\{\rho_d\}$," thus replacing "(1.1) is satisfied by $\{\pi_d\}$." In fact, we describe later (section 2.3) why the two benchmarks lead to identical results in the present discrete-choice context. However, in other contexts such as in the effort example in section 3 there may be no obvious counterpart of (1.1), while the model based on (2.3) is applicable.

2.2 Rationalization

We define what it means for a measure Q over decision rules to rationalize the empirical measure λ. In the extreme case where the analyst knows the distributions over menus this is expressed by (2.3) using the known conditionals $\{\rho_d\}$. One can capture the other extreme of complete ignorance by requiring that (2.3) is satisfied for *some* conditionals $\{\rho_d\}$, restricting them only to reflect certainty that d chooses an element in C_d, that is, $\rho_d \in \Delta(C_d)$ for every d. The associated robustness may be desirable but comes with costs (that we formalize below). First, if "anything goes," then the identified region for any given data λ is large. Second, with such weak maintained assumptions, (almost) every λ can be rationalized by some Q. Consequently, and also because there are situations in which there exists partial information about the menu process, we propose a model that also accommodates intermediate situations.

To model the presence of some information, we assume that, for each d, only distributions ρ_d that lie in the set $\mathcal{R}_d \subset \Delta(C_d)$, determined by the

analyst, are deemed relevant. This leads to the following definition: Say that $Q \in \Delta(\mathcal{D})$ *rationalizes* λ *given* $\{\mathcal{R}_d\}$ *if there exists* $\rho_d \in \mathcal{R}_d$ for all d, such that

$$\lambda(a) = \sum_{d \in \mathcal{D}} Q(d)\, \rho_d(a) \quad \text{for all } a \in X. \tag{2.4}$$

Perfect information is the special case where each $\mathcal{R}_d$ is a singleton. More interesting specifications follow.[6]

Complete ignorance: Let $\mathcal{R}_d = \Delta(C_d)$ for each d as indicated above.

ϵ-contamination: For each d, let $\widehat{\rho}_d \in \Delta(C_d)$ be a focal probability distribution over alternatives, perhaps the analyst's best "point estimate," but one in which she may not have complete confidence. As a reflection of her incomplete confidence she entertains as possible all contaminations of $\widehat{\rho}_d$ of the form

$$\rho_d = (1-\epsilon)\,\widehat{\rho}_d + \epsilon\widetilde{\rho}_d,$$

where $\widetilde{\rho}_d$ is any measure on C_d and where $0 \leq \epsilon \leq 1$ is a parameter to be specified by the analyst. That is, let

$$\begin{aligned} \mathcal{R}_d &= \{\rho_d : \rho_d = (1-\epsilon)\,\widehat{\rho}_d + \epsilon\widetilde{\rho}_d,\ \widetilde{\rho}_d \in \Delta(C_d)\} \\ &= (1-\epsilon)\,\widehat{\rho}_d + \epsilon\Delta(C_d). \end{aligned} \tag{2.5}$$

The extremes $\epsilon = 0, 1$ correspond respectively to the complete confidence and complete ignorance models respectively. Further, it is easy to see that $\mathcal{R}_d$ grows larger in the sense of set inclusion as ϵ increases in $[0, 1]$. This suggests the interpretation of decreasing confidence (or increasing ignorance) as ϵ increases.

The "ϵ-contamination" model has been used frequently in robust statistics (e.g. Huber (1964), Huber and Ronchetti 2009, Wasserman and Kadane 1990), and also in decision theory and its many applications where it is a useful parametric specialization of the set of priors appearing in multiple-priors utility (Gilboa and Schmeidler 1989).

Variation neighborhood: For any p' and p in $\Delta(C_d)$, define

$$\delta_d(p', p) = \sup_{K \subset C_d} \mid p'(K) - p(K) \mid .$$

[6] They are all well-known in both robust statistics and in decision theory. We have borrowed them and their properties described later from Wasserman and Kadane (1990). However, we have not seen the last four used previously in the present context.

Fix a reference/focal measure P_d on C_d and $\epsilon > 0$, and let

$$\mathcal{R}_d = \{p_d \in \Delta(C_d) : \delta_d(p_d, P_d) < \epsilon\}. \tag{2.6}$$

Interval beliefs: Let p_{*d} and p_d^* be measures (not probability measures) on C_d, satisfying

$$p_{*d}(\cdot) \leq p_d^*(\cdot) \text{ and } 0 < p_{*d}(C_d) < 1 < p_d^*(C_d),$$

and define

$$\mathcal{R}_d = \{p_d \in \Delta(C_d) : \ p_{*d}(\cdot) \leq p_d(\cdot) \leq p_d^*(\cdot) \text{ on } C_d\}.$$

In the special case

$$p_{*d} = a_d P_d \text{ and } p_d^* = b_d P_d,$$

where $a_d < 1 < b_d$ and P_d is a (fixed) probability measure on C_d, one obtains

$$\mathcal{R}_d = \{p_d \in \Delta(C_d) : \ a_d P_d \leq p_d \leq b_d P_d\}$$

2-dimensional beliefs: Let P_d^1 and P_d^2 be two distinct probability measures on C_d. The analyst views these measures and all averages (mixtures) as the set of relevant probability laws. Accordingly,

$$\mathcal{R}_d = \{\alpha P_d^1 + (1-\alpha) P_d^2 : 0 \leq \alpha \leq 1\}.$$

In all cases, the identified set is (weakly) smaller than the identified set under complete ignorance. More generally, it shrinks if confidence increases in the sense that each set $\mathcal{R}_d$ shrinks; this happens, for example, if ϵ is reduced in the ϵ-contamination specification or in the variation neighborhood specification. (Similarly, if each set of alternatives C_d shrinks.) It is easy to see also that an increase in confidence shrinks the set of empirical measures λ that can be rationalized by some Q. For example, in the absence of any confidence (complete ignorance), every λ with support in $\cup_d C_d$ can be rationalized by some Q, while in the ϵ-contamination specification λ can be rationalized only if it can be expressed as a mixture $(1-\epsilon)\widehat{\lambda} + \epsilon\widetilde{\lambda}$ where $\widehat{\lambda}$ is rationalizable under complete confidence ($\epsilon = 0$) and $\widetilde{\lambda}$ is rationalizable under zero confidence ($\epsilon = 1$), that is, if

$$\lambda \in (1-\epsilon)\, ch(\{\widehat{\rho}_d\}) + \epsilon\Delta(\cup_d C_d),$$

where *ch* denotes 'convex hull.'

2.3 A characterization

The main question to be addressed is "*which measures Q can rationalize λ given $\{\mathcal{R}_d\}$*?" We provide a comprehensive answer under the assumption that each $\mathcal{R}_d$ is the core of a convex capacity, that is, for each d,[7]

$$\mathcal{R}_d = core\,(\nu_d)\,, \text{ for some } \nu_d \text{ convex.} \tag{2.7}$$

Though limiting, (2.7) is of interest in light of the role of convex capacities in decision theory and in statistics (as mentioned in the introduction); and we note also that it is satisfied by all of the preceding specifications.

Theorem 2.1. *Let $\{\mathcal{R}_d\}$ be such that, for each d, $\mathcal{R}_d = core\,(\nu_d)$ for some convex capacity ν_d on C_d. Then $Q \in \Delta(\mathcal{D})$ rationalizes λ given $\{\mathcal{R}_d\}$ if and only if*

$$\lambda(K) \geq \sum_{d \in \mathcal{D}} Q(d)\,\nu_d(K) \quad \textit{for all } K \subset X. \tag{2.8}$$

In particular, this equivalence applies to the five special cases of $\{\mathcal{R}_d\}$ described above where the corresponding capacities ν_d are given by:

$$\begin{array}{lrcc}
ignorance & \nu_d(K) & = & \mathbf{1}_{C_d \subset K} \\
contamination: & \nu_d(K) & = & (1-\epsilon)\,\widehat{\rho}_d(K \cap C_d) + \epsilon \mathbf{1}_{C_d \subset K} \\
variation\ nbhd & \nu_d(K) & = & \max\{P_d(K \cap C_d) - \epsilon, 0\} \text{ if } C_d \not\subset K \\
 & & & \text{and } = 1 \text{ if } C_d \subset K \\
interval\ beliefs & \nu_d(K) & = & \max\{p_{*d}(K \cap C_d)\,, p_d^*(K \cap C_d) - \beta_d\} \\
 & & & \beta_d = p_d^*(C_d) - 1 \\
2\text{-dim } beliefs & \nu_d(K) & = & \min\{P_d^1(K \cap C_d)\,, P_d^2(K \cap C_d)\}
\end{array}$$

The main message is that the sharp identified set of measures Q is the set of solutions Q to the finite set of linear inequalities (2.8), and constitutes a (convex) polytope. The proof is extremely simple.

Proof: Under the assumption (2.7), rationalizability amounts to the statement that

$$\lambda \in \sum_d Q(d)\ core\,(\nu_d)\,,$$

[7] Since a convex capacity is uniquely determined by its core (see (A.3)), ν_d is necessarily unique. Another point (see the appendix), is that a capacity ν_d on C_d, hence satisfying $\nu_d(C_d) = 1$, can be uniquely extended to a capacity on all of X, just as a probability measure on C_d can be so extended.

while condition (2.8) is the statement that

$$\lambda \in core\left(\sum_{d} Q\left(d\right)\nu_{d}\right).$$

However, by (A.4), the core of the mixture equals the mixture of the cores, which proves the required equivalence.

For the assertions regarding the special cases, one need only show that in each case the indicated capacity ν_d is convex and that it has core equal to the corresponding set $\mathcal{R}_d$. But these are well-known facts (Huber and Strassen 1973, Wasserman and Kadane 1990); for 2-dimensional beliefs, see also Topkis (1998, Lemma 2.6.4). ■

Remark: Theorem 2.1, is, in fact, equivalent to Theorem 4 of Strassen (1965) when specialized as here so that all sets are finite. An important difference is our drastically simpler elementary proof. We view the simpler proof as significant not as a mathematical contribution, but rather because it enhances transparency and accessibility of the theorem which, we believe, may help to expose and promote it as a useful tool for economists. Another value-added over Strassen is our demonstration of the theorem's usefulness as outlined above.[8]. Finally, our proof uses convexity of the ν_ds only to justify applying the mixture-linearity property of their cores, that is, the characterization provided by (2.8) is valid also if convexity is replaced by this mixture-linearity. In this sense, therefore, since Strassen's Theorem 4 assumes convexity, our result is (strictly) more general (albeit given finiteness). Tijs and Branzei (2002) and Bloch and de Clippel (2010) give other assumptions, besides convexity, that imply mixture-linearity of cores.[9] Indeed, it follows from the latter paper that (generically) there exists a partition of the set of all capacities having nonempty cores such that the mixture-linearity property is satisfied if (and only if) all capacities lie in the same equivalence class. The set of convex capacities is one such equivalence class, but there are others and the theorem applies to each of them as well. It remains for future work to determine if any of the other equivalence classes provide alternatives to convexity that are interesting in our setting.

The following discussion provides additional perspective on the theorem and its value-added. Consider first the special case of complete ignorance.

[8]Doval and Eilat (2023) apply Strassen's Theorem 3, in addition to network flow arguments, in their proofs.

[9]They work in the context of cooperative games where capacities are typically not normalized to assign a common fixed value to the universal coalition, and hence they refer to additivity rather than mixture-linearity of the core.

The associated capacities, written more fully, are given by

$$\nu_d(K) = \begin{cases} 1 & C_d \subset K \\ 0 & C_d \not\subset K \end{cases}$$

The epistemic interpretation is that C_d is certain but there is complete ignorance within C_d. The condition (2.8) characterizing rationalizability specializes to the set of inequalities

$$\lambda(K) \geq Q(\{d \in \mathcal{D} : C_d \subset K\}) \quad \text{for all } K \subset X. \tag{2.9}$$

Similar conditions have appeared previously in Barseghyan et al (2021, Theorem 3.1) and in Azrieli and Rehbeck (2023). As indicated in the introduction, the latter addresses different questions, and the former assumes a designated minimum size for menus. The ignorance special case admits alternative proofs. For one, the associated capacities ν_d are belief functions and hence the result for that case can be derived by using random set theory, which is the approach taken by Barsheghyan et al (2021). Alternatively, it follows immediately from the well-known structure of the core of a belief function (Dempster (1967) or Wasserman (1990, Theorem 2.1)). Moreover, these alternatives apply also to the ϵ-contamination specification since its ν_ds are belief functions. However, they do not apply when the ν_ds are convex but not belief functions, such as in the other three special cases or at the level of generality in the theorem.

To illustrate further the greater richness provided by admitting capacities that are convex rather than only belief functions, consider additional specifications that extend those given in the theorem. As defined above the ϵ-contamination specification models an analyst who is concerned that the focal measure may be contaminated by *any* probability measure. In some circumstances, however, only a subset of contaminations may be relevant (in statistics see Berger and Berliner (1986) and Moreno and Cano (1991), and in decision theory see Kopylov (2016)). For example, they might be restricted to lie in a variation neighborhood (2.6), thus leading to the following generalization of (2.5):

$$\mathcal{R}_d = (1-\epsilon)\,\widehat{\rho}_d + \epsilon \mathcal{R}'_d,$$

where $\mathcal{R}'_d \subset \Delta(C_d)$ is defined as in (2.6), for some radius $\epsilon' \neq \epsilon$ about ρ'_d. Then $\mathcal{R}_d$ equals the core of the convex capacity $(1-\epsilon)\,\widehat{\rho}_d + \epsilon\nu'_d$, where ν'_d is the convex capacity whose core is $\mathcal{R}'_d$, and thus is accommodated by the theorem. Similarly for the further generalization whereby

$$\mathcal{R}_d = (1-\epsilon)\,\widehat{\mathcal{R}}_d + \epsilon \mathcal{R}'_d,$$

where $\widehat{\mathcal{R}}_d$ is the variation neighborhood about $\widehat{\rho}_d$ with radius $\widehat{\epsilon}$ (with associated convex capacity $\widehat{\nu}_d$). The interpretation is that there are two focal measures $\widehat{\rho}_d$ and ρ'_d, but each is known (or believed to be valid) only up to small perturbations of size $\widehat{\epsilon}$ and ϵ' respectively, and where the two hypotheses have prior subjective probabilities $(1-\epsilon)$ and ϵ. Note that the theorem applies to this specification because, by the mixture linearity property (A.4),

$$\begin{aligned}\mathcal{R}_d &= (1-\epsilon)\,\widehat{\mathcal{R}}_d + \epsilon\mathcal{R}'_d \\ &= (1-\epsilon)\,core\,(\widehat{\nu}_d) + \epsilon core\,(\nu'_d) \\ &= core\,((1-\epsilon)\,\widehat{\nu}_d + \epsilon\nu'_d) \equiv core\,(\nu_d)\,.\end{aligned}$$

However, none of ν_d, $\widehat{\nu}_d$ and ν'_d is a belief function and therefore this specification is not covered by the random set theory approach in Barseghyan et al (2021).

Application of the theorem requires that when considering whether to adopt a specification of interest for the $\mathcal{R}_d$s one is able to check whether it satisfies (2.7). For the particular specifications addressed in the theorem, the literature has confirmed (2.7). More generally, an important observation is that, given $\{\mathcal{R}_d\}$, then, for each d, there is only one candidate for a suitable capacity ν_d, namely the *lower probability* corresponding to $\mathcal{R}_d$ and defined by

$$\nu_d\,(K) = \inf\{\rho\,(K) : \rho \in \mathcal{R}_d\},\ \text{ for all } K \subset X.$$

In other words, (2.7) is *equivalent* to the assumption that the lower probability capacity is convex and has $\mathcal{R}_d$ as its core. (This follows directly from (A.3).) Convexity of ν_d can be checked, in principle, by using its definition (A.1) or any of its equivalent characterizations (Grabisch (2016, Theorem 3.15), for example). Since $\mathcal{R}_d \subset core\,(\nu_d)$ follows from the definition of lower probability, equality amounts to the requirement that $\mathcal{R}_d$ be sufficiently large in the sense that, for every $\rho \in \Delta\,(C_d)$,

$$\rho\,(K) \geq \nu_d\,(K)\ \text{ for all } K \subset C_d \text{ implies } \rho \in \mathcal{R}_d.^{10}$$

Related is the question what can be done if one drops the assumption (2.7) entirely.[11] In fact, it is straightforward to show that the counterpart

[10] Alternatively, given convexity, one can compute the cores by using the greedy algorithm Ichiishi (1981), or the algorithm in Chambers and Melkonyan (2005) that uses information about willingness to buy or sell and thus may help the analyst to calibrate parameters like ϵ.

[11] If the sets $\{C_d\}$ are disjoint, then, for any $\{\mathcal{R}_d\}$, there is point identification - λ is rationalized by the *unique* measure Q given by $Q\,(d) = \lambda\,(C_d)$ for all d.

of (2.8) given below is *necessary* for rationalizability given $\{\mathcal{R}_d\}$: If $Q \in \Delta(\mathcal{D})$ rationalizes λ, then, for the lower probability capacity ν_d, (using $\mathcal{R}_d \subset core(\nu_d)$ and (A.5)),

$$\begin{aligned} \lambda \in \sum_d Q(d)\,\mathcal{R}_d &\subset \sum_d Q(d)\,core(\nu_d) \\ &\subset core\left(\sum_d Q(d)\,\nu_d\right) \implies \\ \lambda(K) \geq \sum_d Q(d)\,\nu_d(K \cap C_d) &\quad \text{for all } K \subset X. \end{aligned}$$

The theorem is relevant also to a rationalizability notion such as (1.1), suitably modified, where menus appear explicitly. Modify (1.1) by allowing π_d, for each d, to vary over a set $\Pi_d \subset \Delta(\mathcal{A})$, where Π_d is determined by the analyst. (For example, $\Pi_d = \Delta(\mathcal{A})$ would model complete ignorance about menus.) Say that $Q \in \Delta(\mathcal{D})$ *menu-rationalizes* λ *given* $\{\Pi_d\}$ *if there exists* $\pi_d \in \Pi_d$ for all d, such that

$$\lambda(a) = \sum_{d \in \mathcal{D}} \sum_{A \in \mathcal{A}} Q(d)\,\pi_d(A)\,\mathbf{1}_{d(A)=a}, \quad \text{for all } a \in X. \tag{2.10}$$

Each π_d induces a distribution ρ_d on C_d as in (2.2); let $\mathcal{R}_d$ be the set of all such distributions ρ_d as π_d varies over Π_d. Then it is immediate that menu-rationalization of λ given $\{\Pi_d\}$ implies rationalization (2.4) given $\{\mathcal{R}_d\}$. Moreover, if Π_d is the core of a convex capacity on $\mathcal{A}$, then $\mathcal{R}_d$ is the core of a convex capacity ν_d on C_d (see appendix). Hence (2.8) is necessary for menu-rationalizability by Q. To prove sufficiency, suppose that Q rationalizes λ given $\{\mathcal{R}_d = core(\nu_d)\}$ and define $\{\Pi_d\} \subset \Delta(\mathcal{A})$ as follows: for each d and $\rho_d \in \mathcal{R}_d$, and for each $a \in C_d$, select one menu $A_{a,d}$ satisfying $d(A_{a,d}) = a$, and define

$$\pi_d(A) = \begin{cases} \rho_d(a) & A = A_{a,d} \\ 0 & A \neq A_{a,d} \end{cases}$$

Then π_d is a probability measure because

$$\sum_{A \in \mathcal{A}} \pi_d(A) = \sum_{a \in C_d} \rho_d(a) = \rho_d(C_d) = 1.$$

Let Π_d be the set of all such π_ds as ρ_d varies over $\mathcal{R}_d$. Then Q menu-rationalizes λ given $\{\Pi_d\}$ because

$$\pi_d(\{A \in \mathcal{A} : d(A) = a\}) = \rho_d(a) \implies$$

$$\sum_d Q(d) \sum_{A \in \mathcal{A}} \pi_d(A) \mathbf{1}_{d(A)=a} = \sum_d Q(d) \rho_d(a) = \lambda(a).$$

(Note that under complete ignorance ($\mathcal{R}_d = \Delta(C_d)$ or $\Pi_d = \Delta(\mathcal{A})$), the above proves the equivalence of the two notions of rationalizability.)

3 Concluding illustrations of scope

We conclude by describing two additional settings where our theorem can be applied to characterize the robust identification of heterogeneity. The first is an instance of our choice model where decision rules are not based on preference maximization. The second, which is not directly connected to choice, concerns the identification of (unobservable) heterogeneous effort.

3.1 Satisficing

There is a population of satisficing decision makers whose aspiration thresholds may differ. They each choose an alternative from the set X and they agree that the value of alternatives is described by $v : X \mapsto \mathbb{R}$. However, individuals differ in two respects. First, aspiration thresholds differ; the set of distinct thresholds is $\mathcal{V}$. Second, individuals differ in the order in which they consider alternatives (this may be a subjective choice or exogenously imposed). Each sequential procedure follows a strict total order $>$ on X: the individual chooses the $>$-first alternative with a value at least as large as her threshold $v \in \mathcal{V}$, and if there are no such "satisfactory" alternatives then she chooses the $>$-last element in X. The empirical frequency of choices λ is observed, but both aspiration levels and orders $>$ are unobserved. Theorem 2.1, suitably reinterpreted, can be used to partially identify the distribution of aspiration levels while respecting limited knowledge (or complete ignorance) of the distribution of orders $>$.

Similar applications can be made to other problems of choice with frames (Salant and Rubinstein 2008) where frames vary across individuals and are unobserved by the analyst.

3.2 Identifying effort

Consider a population of workers with common observable characteristics (e.g. education and experience) who work independently. Each produces a homogeneous output in quantity represented by an element of X.[12] The

[12] To make clear the connection to the main choice model, we use the same symbols, though with different interpretations.

empirical frequency distribution of outputs is given by $\lambda \in \Delta(X)$. Heterogeneity in output is attributed to differences in unobservable characteristics. The first unobservable is effort - there are finitely many effort levels $d \in \mathcal{D}$. The other unobservable is "everything else." The analyst may not be able to describe these other factors precisely, or even at all. However, she takes a stand on the set of their possible output consequences. Formally, for each effort d, denote by $C_d \subset X$ the set of outputs possible given the effort level and given what may ensue from "everything else." The analyst specifies the sets C_d, but is ignorant about likelihoods *within* these sets.[13]

With this reinterpretation, rationalizability of λ is well-defined, and Theorem 2.1 can be applied to yield the (computationally tractable) sharp identified set of measures Q over effort levels.

The rationalizability notion (2.10) could also be accommodated by introducing a parameter $\theta \in \Theta$ to represent "everything else," and, for each d, a production function f such that the pair (d, θ) yields output $f(d, \theta)$, and $C_d = \{f(d, \theta) : \theta \in \Theta\}$. Ignorance about Θ would be captured by admitting any distribution π_d over Θ in the counterpart of definition (2.10). (Roughly, Θ would play the role of the set of menus $\mathcal{A}$ above.) However, such a formulation involving production functions f and probability distributions over Θ is arguably problematic in situations where the analyst cannot even conceive of what is included in "everything else."

A Appendix: Basic facts about capacities

For any finite set X, ν is a *capacity* on X if $\nu : 2^X \to [0,1]$, $\nu(\varnothing) = 0$, $\nu(X) = 1$ and $\nu(K') \geq \nu(K)$ whenever K' is a superset of K. ν is *convex* if, for all subsets K' and K,

$$\nu(K' \cup K) + \nu(K' \cap K) \geq \nu(K') + \nu(K). \tag{A.1}$$

ν is a *belief function* if, for all n, and for all subsets $K_1, ..., K_n$,

$$\nu\left(\cup_{j=1}^{n} K_j\right) \geq \sum_{\varnothing \neq J \subset \{1,...,n\}} (-1)^{|J|+1} \nu\left(\cap_{j \in J} K_j\right). \tag{A.2}$$

If one restricts n to be 2, then one obtains the condition defining convexity. Hence every belief function is convex. (Convexity is sometimes referred to as monotonicity of order 2 while (A.2) is called infinite or total monotonicity.)

[13] The assumption that C_d can be specified even though "everything else" is poorly understood brings to mind Maskin and Tirole (1999) who argue that optimal contracts survive even with unforeseen *contingencies* when agents can forecast future *payoffs*.

A more transparent and equivalent definition of a belief function is that ν is induced by a random set.[14]

Let C be a subset of X and ν a capacity on C (hence $\nu(C) = 1$). Then ν can be viewed also as a capacity on X by identifying ν with the capacity ν' on X defined by

$$\nu'(K) = \nu(K \cap C) \text{ for all } K \subset X.$$

Further, ν' is convex if and only if ν is convex. We often identify ν and ν' and do not distinguish them notationally.

For any capacity ν on X, its core is the set of all dominating probability measures, that is,

$$core(\nu) = \{p \in \Delta(X) : p(K) \geq \nu(K) \text{ for all } K \subset X\}.$$

If ν is convex, then its core is nonempty and ν can be recovered from its core as its lower bound or envelope:

$$\nu(K) = \min\{p(K) : p \in core(\nu)\}. \tag{A.3}$$

If $\nu = p$ is a probability measure, then it is convex and $core(\nu) = \{p\}$.

If ν and ν' are two convex capacities on X, and if $0 \leq \alpha \leq 1$, then the mixture $\alpha\nu + (1-\alpha)\nu'$ is also a convex capacity and its core satisfies

$$core(\alpha\nu + (1-\alpha)\nu') = \alpha core(\nu) + (1-\alpha) core(\nu'). \tag{A.4}$$

(See Danilov and Koshevoy (2000, p. 9) or Grabisch (2016, p. 156).) This "mixture linearity" of the core is the key property that we exploit to prove our theorem. Elsewhere, we also make use of the following weaker, and elementary, property that applies to any (not necessarily convex) capacities

$$core(\alpha\nu + (1-\alpha)\nu') \supset \alpha core(\nu) + (1-\alpha) core(\nu'). \tag{A.5}$$

Let ψ be a convex capacity on $\mathcal{A}$, $\Pi = core(\psi)$ and $d : \mathcal{A} \longrightarrow X$. Define the (convex) set $\mathcal{R}$ of all measures $\rho_\pi \in \Delta(X)$, where

$$\rho_\pi(K) = \pi\left(d^{-1}(K)\right), \text{ for all } K \subset X,$$

and define the set function ν on X by

$$\nu(K) = \psi\left(d^{-1}(K)\right), \text{ for all } K \subset X.$$

[14] Dempster (1967) and Nguyen (1978) are two early references describing the connection of random sets to belief functions. See also Nguyen (2006).

Then ν is a convex capacity and $\mathcal{R} = core(\nu)$. (Convexity follows from verifying (A.1), and $\mathcal{R} \subset core(\nu)$ is immediate. Let $\{K_j\}$ be any chain of subsets of X. Then $\{d^{-1}(K_j)\}$ is a chain in $\mathcal{A}$. Since ψ is convex, there exists $\pi^* \in core(\psi) = \Pi$ such that $\pi^*(d^{-1}(K_j)) = \psi(d^{-1}(K_j))$, for all j (Choquet 1953). Thus $\rho_{\pi^*}(K_j) = \nu(K_j)$ for all j. Apply Grabisch (2016, Theorem 3.15) to conclude that $\mathcal{R} = core(\nu)$.)

Acknowledgements: We are extremely indebted to Rohan Dutta who was instrumental from conception through execution. We thank also Max Amarante and participants at the McGill applied micro breakfast for suggestions.
Funding: This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

References

[1] Abaluck J. and Adams-Prassl A. What do consumers consider before they choose? identification from asymmetric demand responses. *Quart. J. Econ.* (2021) 136, 1611-1663.

[2] Azrieli Y. and Rehbeck J. Marginal stochastic choice. Mimeo 2023.

[3] Barseghyan L., Coughlin M., Molinari F., and Teitelbaum J.C. Heterogenous choice sets and preferences. *Econometrica* (2021) 89(5), 2015-2048.

[4] Berger J. and Berliner M. Robust Bayes and empirical analysis with epsilon-contaminated priors. *Ann. Statist.* (1986) 14, 461-486.

[5] Bloch F. and de Clippel G. Cores of combined games. *J. Econ. Theory* (2010) 145, 2424-2434.

[6] Camerer C. Individual decision-making, pp. 587-704 in *The Handbook of Experimental Economics 1*, J. Kagel and A. Roth eds. Princeton U. Press, 1995.

[7] Caplin A., Dean M., and Leahy J. Rational inattention, optimal consideration sets, and stochastic choice. *Rev. Econ. Stud.* (2019) 86, 1061-1094.

[8] Cattaneo M.D., Ma X., Masatlioglu Y., and Suleymanov E. A random attention model. *J. Pol. Econ.* (2020) 128, 2796-2836.

[9] Chambers R.G. and Melkonyan T.A. Eliciting the core of a supermodular capacity. *Econ. Theory* (2005) 26, 203-209.

[10] Choquet G. Theory of capacities. *Annales de l'institut Fourier* (1954) 5, 131-295.

[11] Danilov V.I. and Koshevoy G.A. Cores of cooperative games, superdifferentials of functions, and the Minkowski difference of sets. *J. Math. Anal. Appl.* (2000) 247, 1-14.

[12] Dardanoni V., Manzini P., Mariotti M. and Tyson C.J. Inferring cognitive heterogeneity from aggregate choices. *Econometrica* (2020) 88(3), 1269-1296.

[13] Dempster A.P. A generalization of Bayesian inference. *J. Royal Statist. Soc.* (B) (1968) 30, 205-247.

[14] Dempster A.P. Upper and lower probabilities induced by multivalued mappings. *Ann. Math. Statist.* (1967) 38, 325-335.

[15] Doval L. and Eilat R. The core of Bayesian persuasion. arxiv 2023.

[16] Filiz-Ozbay E. and Masatlioglu Y. Progressive random choice. *J. Pol. Econ.* (2023) 131, 716-750.

[17] Galichon A. and Henry M. Set identification in models with multiple equilibria. *Rev. Econ. Stud.* (2011) 78, 1264-1298.

[18] Goeree M.S. Limited information and advertising in the U.S. personal computer industry. *Econometrica* (2008) 76, 1017-1074.

[19] Grabisch M. *Set Functions, Games and Capacities in Decision Making.* Springer, 2016.

[20] Huber P.J. Robust estimation of a location parameter. *Ann. Math. Statist.* (1964) 35(1), 73-101.

[21] Huber P.J. and Ronchetti E.M. *Robust Statistics.* 2nd ed. Wiley, 2009.

[22] Huber P.J. and Strassen V. Minimax tests and the Neyman-Pearson lemma for capacities. *Ann. Statist.* (1973) 1, 251-263.

[23] Ichiishi T. Super-modularity: applications to convex games and to the greedy algorithm for LP. *J. Econ. Theory* (1981) 25, 283-286.

[24] Kopylov I. Subjective probability, confidence and Bayesian updating. *Econ. Theory (2016) 62, 635-658.*

[25] Lu Z. Estimating multinomial choice models with unobserved choice sets. *J. Econometrics* (2022) 226, 368-398.

[26] Manski C. The structure of random utility models. *Theory and Decision* (1977) 8(3), 229-254.

[27] Manzini P. and Mariotti M. Stochastic choice and consideration sets. *Econometrica* (2014) 82, 1153-1176.

[28] Masatlioglu Y., Nakajima D., and Ozbay E. Revealed attention. *Amer. Econ. Rev.* (2012) 102, 2183-2205.

[29] Maskin E. and Tirole J. Unforeseen contingencies and incomplete contracts. *Rev. Econ. Stud.* (1999) 66, 83-114.

[30] McFadden, D.L. Conditional logit analysis of qualitative choice behavior, pp. 105-142 in *Frontiers of Econometrics,* ed. P. Zarembka. New York: Academic Press 1974.

[31] Moreno E. and Cano J. Robust Bayesian analysis with epsilon-contaminations partially known. *J. Royal Statist. Soc.* (B) (1991) 53, 143-155.

[32] Nguyen H.T. On random sets and belief functions. *J. Math. Anal. Appl.* (1978) 68, 531-542.

[33] Nguyen H.T. *An Introduction to Random Sets.* Chapman and Hall, 2006.

[34] Salant Y. and Rubinstein A. (A,f): Choice with frames. *Rev. Econ. Stud.* (2008) 75(4), 1287-1296.

[35] Schmeidler D. Subjective probability and expected utility without additivity. *Econometrica* (1989) 57(3), 571-587.

[36] Shafer G. *Mathematical Theory of Evidence.* Princeton U. Press, 1976.

[37] Strassen V. The existence of probability measures with given marginals. *Ann. Math. Statist.* (1965) 36, 81-95.

[38] Tijs S. and Branzei R. Additive stable solutions on perfect cones of cooperative games. *Int. J. Game Theory* (2002) 31, 469-474.

[39] Topkis D.M. *Supermodularity and Complementarity.* Princeton U. Press, 1998.

[40] Wasserman L.A. Prior envelopes based on belief functions. *Ann. Statist.* (1990) 18(1), 454-464.

[41] Wasserman L.A. and Kadane J.B. Bayes' theorem for Choquet capacities. *Ann. Math. Statist.* (1990) 18, 1328-1339.