Available online at www.sciencedirect.com

ScienceDirect

Journal of Economic Theory 168 (2017) 329–371

JOURNAL OF Economic Theory

www.elsevier.com/locate/jet

Interdependent preferences and strategic distinguishability

Dirk Bergemann [a], Stephen Morris [b], Satoru Takahashi [c]

[a] *Yale University, USA*
[b] *Princeton University, USA*
[c] *National University of Singapore, Singapore*

Received 18 July 2014; final version received 28 December 2016; accepted 3 January 2017

Available online 6 January 2017

Abstract

We study agents whose expected utility preferences are interdependent for informational or psychological reasons. We characterize when two types can be "strategically distinguished" in the sense that they are guaranteed to behave differently in some finite mechanism. We show that two types are strategically distinguishable if and only if they have different hierarchies of interdependent preferences. The same characterization applies for rationalizability, equilibrium, and any interim solution concept in between. Our results generalize and unify results of Abreu and Matsushima (1992), who characterize strategic distinguishability on fixed finite type spaces, and Dekel et al. (2006, 2007), who characterize strategic distinguishability without interdependent preferences.

JEL classification: C79; D82; D83

Keywords: Interdependent preferences; Higher order preferences; Hierarchy of preferences; Strategic distinguishability

1. Introduction

Consider a setting where agents have preferences over lotteries on a finite set of outcomes, and their preferences are interdependent. Such interdependence may arise for *informational* reasons:

E-mail addresses: dirk.bergemann@yale.edu (D. Bergemann), smorris@princeton.edu (S. Morris), ecsst@nus.edu.sg (S. Takahashi).

http://dx.doi.org/10.1016/j.jet.2017.01.002
0022-0531/

an agent believes that another agent's preferences encode information that she thinks is relevant for her own payoffs. Or interdependence may exist for *psychological* reasons: an agent may be more altruistic to an agent who she thinks is more altruistic. Whatever the reason for the interdependence, the standard approach to modeling such interdependence is to let each agent have a set of possible "types," where each type has a belief about the other agents' types, and each agent has a utility function specifying her utility over outcomes given each profile of agents' types. A difficulty with this approach is that when types are described in this "implicit", self-referential, way, the operational content of such type descriptions is not clear. In particular, the distinction between informational and psychological interdependence may not have observable implications. In this paper, we provide a unified treatment of interdependent preferences and characterize when two such types can be "strategically distinguished" in the sense that they are guaranteed to behave differently in some finite mechanism.

To present our characterization of strategic distinguishability in the simplest possible environment, we assume that each agent's preference can be represented by expected utilities that depend on the other agents' type profile. Furthermore, we assume (1) uniform ranking: for each agent, there are two outcomes, a "good outcome" and a "bad outcome," such that he strictly prefers the former to the latter given any profile of the other agents' types, and (2) bounded utilities: his utility indices lie in a prefixed bounded set when we normalize the utility of the "good outcome" to 1 and the "bad outcome" to 0. These assumptions are with loss of generality.[1] But because of these assumptions, we can find finitely many "extreme" utility indices such that each utility index in the bounded set can be expressed as a convex combination of those extreme utility indices uniquely. Fixing and interpreting such extreme utility indices as his "private states," we express any of his utility indices as a probability distribution over the private states, and interdependent preferences by a type space based on the profile of these private states. Now each type has a hierarchy of beliefs about the private states à la Mertens and Zamir (1985). An agent's hierarchy of beliefs about the private states represents a "hierarchy of interdependent preferences": a first order belief represents a preference over lotteries, which we call a "first order preference"; a second order belief represents a preference over (Anscombe–Aumann) acts over the opponents' first order preferences, which we call a "second order preference", and so on. With such a representation, our main result states that two types are strategically distinguishable if and only if they have different hierarchies of beliefs about the private states and thus if they have different hierarchies of interdependent preferences. The same characterization applies for interim correlated rationalizability, equilibrium, and any interim solution concept in between (Theorems 1 and 2).

The question of strategic distinguishability is due to Abreu and Matsushima (1992) (AM). AM characterize (full) virtual Bayesian implementability of social choice functions for a finite type space under the solution concept of iterated deletion of strictly dominated strategies (as well as equilibrium).[2] A necessary condition is a "measurability" condition that, in the language of

[1] The uniform ranking condition is weaker than the *economic condition* maintained in much of the implementation literature, see Palfrey and Srivastava (1989) or Jackson (1991). The economic condition requires that a uniform strict ranking over two alternatives by agent i is reversed by the ranking of another agent j. It is called an economic condition because – in an exchange economy – giving one agent a strictly larger, and thus strictly preferred, bundle, requires giving another agent a smaller, less preferred bundle. Nonetheless, uniform ranking remains a strong assumption and we will discuss below how it can be relaxed.

[2] A technical difference between AM and our paper is that we assume the outcome set to be finite whereas AM consider a more general environment and allow for all simple lotteries, i.e., lotteries with finite support, over an arbitrary, possibly infinite, outcome set.

this paper, requires that the social choice function gives the same outcome to profiles of types with the same hierarchies of interdependent preferences. Lemma 2 of AM shows that two types in a given finite type space are strategically distinguishable if and only if they have different hierarchies of interdependent preferences. Subject to our uniform ranking assumption, we make two contributions relative to AM: first, we do not require a finite type space. Second, we show that for every fixed pair of distinct hierarchies of interdependent preferences, there is a single finite mechanism that will strategically distinguish types with those hierarchies of interdependent preferences across all type spaces, finite or infinite.

We use beliefs (and hierarchies of beliefs) over private states to represent preferences (and hierarchies of interdependent preferences) over outcomes, and – for our main results – mechanisms depend only on actions. It is straightforward to extend our results to allow for additional *external states*: states that mechanisms can be made conditional on (Section 5.1). Now strategic distinguishability is characterized by hierarchies of beliefs over the private states as well as these external states. Dekel et al. (2006, 2007) (DFM) consider a setting where agents have beliefs and higher order beliefs about external states, and show that two types have disjoint sets of rationalizable actions in some finite game with external-state-dependent payoffs if and only if they have different hierarchies of beliefs about the external states (Dekel et al., 2006, Lemma 4) and Dekel et al. (2007, Proposition 1 and Corollary 2). Our results then reduce to DFM's result when each agent has a single private state, i.e., when there is common certainty of preferences and – in particular – no interdependence of preferences (Section 5.2). Indeed, DFM's result would go through even if we restricted attention to special classes of games such as zero sum games or common interest games.[3]

Thus our results can be seen as a unification and extension of the results of AM and DFM. Like DFM and unlike AM, we use an explicit description of types (independent of the type space they belong to) and can distinguish types on arbitrary type spaces. Like AM and unlike DFM, we distinguish between types with different hierarchies of interdependent preferences and not just types with different hierarchies of beliefs about external states, and thus we are more constrained in the set of strategic settings we can confront agents with.

Our main result requires an innovation in the proof strategy. In order to strategically distinguish two types with distinct hierarchies of interdependent preferences, we construct a finite mechanism in which agents are asked to report first finite orders of preferences. The mechanism randomizes over which "component" of the mechanism is used to select the outcome. For each agent i and each order n, there is an (i,n)th component of the mechanism designed so that agent i has an incentive to truthfully report her nth order preference if other agents have truthfully reported their $(n-1)$th and lower order preferences. A potential difficulty with this proof strategy is that agent i's report of an nth order preference is an input not only into the (i,n)th component of the mechanism, but also into the $(j,n+1)$th components, i.e., the components giving each other agent j an incentive to truthfully report his $(n+1)$th and higher order preferences. AM dealt with this difficulty by exploiting finiteness, and having the probability of (j,m)th components, for all $m \geq n+1$, occur with much smaller probability than component (i,n). DFM can choose payoffs so that, for each agent i, the (i,n)th components (for all n) giving agent i an incentive to report her preferences truthfully have no implications for other agents' incentives. Neither trick is available in our setting, as we have arbitrary type spaces and agents' preferences over outcomes may be arbitrarily linked. Instead, we develop a *robust scoring rule* that not only

[3] Gossner and Mertens (2001) suggested such a result for zero sum games.

gives an agent an incentive to report her nth order preferences truthfully if others report their $(n-1)$th and lower order preferences truthfully, but also gives the agent an incentive to report her nth order preferences *approximately* truthfully if others report their $(n-1)$th and lower order preferences approximately truthfully. This enables us to design a mechanism where the error size at each of a finite number of orders can be simultaneously controlled.[4]

As an extension, we relax the restriction on preferences. Namely, we replace the uniform ranking assumption by the "no complete indifference" assumption, i.e., we assume that no type is completely indifferent over outcomes, but the "good outcome" and the "bad outcome" may depend on own types as well as the opponents' types. Without the uniform ranking assumption, we may not be able to define private states (i.e., extremal preferences) in a meaningful way, not to mention a hierarchy of beliefs about the private states, as not all conditional preferences given type profiles may be represented by a convex combination of finitely many extreme utility indices. But we can still make sense of a hierarchy of interdependent preferences as a sequence of preferences, the first term denoting a preference over lotteries (first order preference), the second term denoting a preference over acts over the opponents' first order preferences (second order preference), and so on. With this terminology, we can extend the main result and show that under the assumptions of *no complete indifference* and *bounded utility*, two types are strategically distinguishable if and only if they have different hierarchies of interdependent preferences. Moreover, the bounded utility assumption may also be relaxed to a certain "λ-continuity" assumption.[5] Similarly to the previous formulation, this characterization holds for a suitably defined version of rationalizability, equilibrium, and any interim solution concept in between.

With this extension, our results imply Lemma 2 of AM but not vice versa, as any given finite type space satisfies the bounded utility assumption with a sufficiently large bound, but the mechanism constructed by AM can only distinguish a pair of distinct hierarchies of interdependent preferences if, in addition, we fix the finite type space to which they belong.

The paper is organized as follows. In Section 2, we discuss an example where each agent's conditional preferences over lotteries, given the opponents' type profiles, are parameterized by a single number in the interval $[0, 1]$, and use the example to motivate why it is hierarchies of beliefs about extreme points of possible preferences which characterize strategic distinguishability, and point out why alternative ways of representing interdependent preferences – implicitly or explicitly considered in the literature – are either not "rich enough" (since they do not describe possible interdependent preference types of interest) or they are not "tight" (separating types that are not strategically distinguishable). In Section 3, we formally introduce our model under the uniform ranking and bounded utility assumptions. In Section 4, under these assumptions, we show that strategic distinguishability is characterized by hierarchies of interdependent preferences represented by hierarchies of beliefs about private states. In Section 5, we discuss two extensions, how to incorporate external states, and how to replace the uniform ranking assumption by the no complete indifference assumption. In Section 6, we discuss further connections to the literature.

[4] A related issue arises in the work of Chambers and Lambert (2014), where the problem of eliciting dynamic (rather than interactive) beliefs is studied.

[5] Without the bounded utility assumption or the λ-continuity assumption, strategic distinguishability would no longer be characterized by hierarchies of interdependent preferences (Proposition 5).

2. Strategically distinguishable hierarchies

Before introducing our formal framework, we give a motivating example to illustrate which types can be strategically distinguished. Consider two "conditionally altruistic agents."[6] Each agent may care about the other's private consumption, so that she is *altruistic*. But an agent may also be more altruistic if she thinks that the other agent is more altruistic, in which case she is *conditionally altruistic*. Higher order conditional altruism is also possible. More concretely, suppose that a prize is being allocated to either of the two agents. There is a probability $r_i \in [0, 1]$ such that agent i is indifferent between the other agent getting the object for sure and getting the object herself with probability r_i. Thus, r_i is an index of agent i's altruism. Conditional altruism corresponds to having a higher altruism index when the other agent has a higher index.

Agent i's interdependent preference type will have the following hierarchical description:[7]

1. A first order preference given by an altruism index r_i describing the agent's preference over lotteries over outcomes.
2. A second order preference over Anscombe–Aumann acts giving outcomes as functions of the other agent's altruism index.
3. A third order preference over Anscombe–Aumann acts giving outcomes as functions of the other agent's second order preference.
4. And so on....

As noted in the introduction, if the altruism index $r_i \in [0, 1]$ is identified with probability distributions over states 0 and 1, then second order preferences correspond to probability distributions over $\{0, 1\} \times \Delta(\{0, 1\})$, which is isomorphic to $\{0, 1\} \times [0, 1]$; a third order preference corresponds to a probability distribution over $\{0, 1\} \times \Delta(\{0, 1\} \times [0, 1])$; and so on. Following standard results in the belief hierarchy literature, we can identify these hierarchies of interdependent preferences with a set T^* consisting of the universal set of belief hierarchies, satisfying coherence and common certainty of coherence, about the extreme points of own altruism indices, where the set T^* satisfies the homeomorphism

$$T^* \cong \Delta(\{0, 1\} \times T^*).$$

Thus, each agent's type is uniquely identified with a belief over $\{0, 1\} \times T^*$. The interpretation is that the marginal belief on T^* corresponds to the agent's belief over the other agent's type; and the conditional probability of state 1 given the other agent's type corresponds to the agent's expected value of the altruism index, conditional on the other agent's type.

Our main result will be that T^* describes the interdependent preference hierarchies that can be strategically distinguished. To provide intuition for our main result, it is useful to discuss three alternative descriptions of interdependent preference hierarchies that have been (implicitly or explicitly) proposed in the literature. We show why each one of the descriptions is either not rich enough – in the sense that it does not include possible interdependent preference hierarchies – or is not tight, in the sense that it labels types differently even if they are not strategically distinguishable.

[6] This discussion follows Levine (1998) and Gul and Pesendorfer (2016).

[7] We will formalize the notion of interdependent preference hierarchies in Section 3.4 by means of beliefs and higher order beliefs over "private states" and in Section 5.3 more directly.

First, we could say (as before) that a first order preference corresponds to an altruism index $r_i \in [0, 1]$. But then we could say that a second order preference is given by an agent's altruism index, and a belief over the other agent's altruism index, and so by an element of $[0, 1] \times \Delta([0, 1])$. Iterating in this way, we would get a "*private values* (PV) *universal belief space*," T_{PV}, satisfying the homeomorphism

$$T_{PV} \cong [0, 1] \times \Delta(T_{PV}).$$

Thus, a type is identified with an altruism index and a belief over the other agent's altruism index. The space T_{PV} clearly allows for all first order preferences. But it does not allow second order preferences to depend on the other agent's first order preference, and hence rules out the interdependence we are trying to capture. T_{PV} is thus not rich enough, although it is tight in that every pair of distinct types is strategically distinguishable.

Second, we could identify types with belief hierarchies, satisfying coherence and common certainty of coherence, about both agents' altruism indices. A first order preference would now be an element of $\Delta\left([0, 1]^2\right)$. A second order preference would be an element of $\Delta([0, 1]^2 \times \Delta([0, 1]^2))$; and so on. Iterating in this way, we would get the "*payoff* (P) *universal type space*," satisfying the homeomorphism

$$T_P \cong \Delta\left([0, 1]^2 \times T_P\right).$$

Now a type is identified with a belief over both agents' altruism indices and the type of the other agent. The space T_P is rich enough to allow for all interdependent preferences we are interested in, but it is not tight as it labels types differently even if they are not strategically distinguishable. For example, it labels a type of agent i who is sure that his altruism index is $1/2$ differently from another type having a $50/50$ belief about whether his altruism index is 0 or 1. It also labels differently types of agent i who have various beliefs about agent j's altruism index r_j, but all of whom are sure that j is sure that j is "truly selfish" (i.e., $r_j = 0$), and j will never behave in an altruistic way. Note that what matters for agent i's behavior in strategic settings is not what agent i believes about r_j, but what agent i believes about agent j's belief about r_j.

Third, we could identify types with beliefs and higher order beliefs about a large set of "payoff types" that describe interdependent preferences (without beliefs). An agent knows his own payoff type but may not know the other agent's payoff type. Thus, suppose that we have a set Ψ of possible payoff types for each agent and let $r\left(\psi, \psi'\right) \in [0, 1]$ specify an agent's altruism index when he has payoff type ψ and the other agent has payoff type ψ', so $r : \Psi^2 \to [0, 1]$. Now an agent's first order preference will consist of a payoff type ψ. His second order preference will consist of a payoff type ψ and a belief over the other agent's payoff type, and will thus be an element of $\Psi \times \Delta(\Psi)$. And so on. Call the set of all such hierarchies the "*interdependent payoff* (IP) *universal belief space*," T_{IP}; it will satisfy the homeomorphism

$$T_{IP} \cong \Psi \times \Delta(T_{IP}).$$

So a type now corresponds to a payoff type and a belief over the other agent's type. Since we assumed that agents knew their own "payoff types," this is simply the private values universal type space defined over Ψ instead of $[0, 1]$ as we did for T_{PV} or $\{0, 1\}$ as we did for T^*. This modeling approach follows a standard practice in the literature of treating payoff interdependence and higher order beliefs separately, and is widely used in the mechanism design literature, either implicitly or explicitly. It is implicit in Dasgupta and Maskin (2000), who introduce "types" which determine players' interdependent values and then consider ways of implementing the

efficient outcome that do not depend on beliefs. It is explicit in the work of two of us on robust mechanism design (Bergemann and Morris (2005, 2009a, 2012) for a collection of related work), where we assumed a space of possible "payoff types," and allow all beliefs and higher order beliefs about those payoff types.

The payoff type spaces in Dasgupta and Maskin (2000) and Bergemann and Morris (2012) are not intended to be "universal." Gul and Pesendorfer (2016) constructed a universal type space of interdependent preferences, abstracting from any belief structure. In particular, they identify a maximal set of interdependent payoff types which captures all distinctions that can be expressed in a natural language. When they consider applications of their universal type space to incomplete information settings, they treat incomplete information separately and thus implicitly allow all beliefs and higher order beliefs over their universal payoff space.

Similarly to the space T_P, the space T_{IP} is rich enough to express all interdependent preferences if the underlying payoff type space is large enough, but it is then not tight. An agent's type in T_{IP} specifies what his payoff parameter would be given the other agent's payoff type that he attaches probability zero to. Thus, it contains information that the agent (subjectively) regards as counterfactual. While there might be purposes for which we want a language to express this information, as discussed in Gul and Pesendorfer (2016), such distinctions will not be strategically distinguishable in our sense. Concretely, suppose that there were two payoff types, an "unconditionally altruistic" type ψ and a "conditionally altruistic" type ψ' with $r(\psi, \psi) = r(\psi, \psi') = r(\psi', \psi) = 1$ and $r(\psi', \psi') = 0$, and compare (i) a type of an agent who prefers sharing with the other agent because he is conditionally altruistic and is sure that the other agent is unconditionally altruistic; and (ii) another type who prefers sharing with the other agent because he is unconditionally altruistic. These types will not be strategically distinguishable from each other, but will correspond to different types in T_{IP}.

Before moving to our main model, let us consider two broader interpretations of the example. First, we could allow the altruism index to be in the interval $[-B, B]$, with the interpretation that $r_i > 1$ corresponds to a super-altruistic agent who prefers the other agent to get the object to getting it himself; and $r_i < 0$ corresponds to a spiteful agent who would prefer that no one got the object to the other agent getting the object. This case is discussed in Section 3.2.

Second, we could allow the parameter $r_i \in [0, 1]$ to have very different interpretations from conditional altruism. For example, suppose that there were three outcomes, bad, intermediate and good, and an agent always strictly preferred the good outcome to the bad outcome and r_i represented agent i's von Neumann–Morgenstern utility index of the intermediate outcome. Or suppose that $r_i \in [0, 1]$ corresponded to agent i's willingness to pay for an object in terms of a numeraire good. The discussion above applies unchanged to these alternative interpretations of the payoff relevant parameter. The latter interpretation corresponds to the leading example in the mechanism design work of Dasgupta and Maskin (2000) and Bergemann and Morris (2012).

3. Model

3.1. Conventions

We record some terminological conventions used throughout the paper. A finite set is endowed with the discrete topology. A countable set is endowed with the discrete σ-algebra. A compact metric space is endowed with the Borel σ-algebra. A countable product $\prod_{n=0}^{\infty} X_n$ of measurable spaces $(X_n)_{n=0}^{\infty}$ is endowed with the product σ-algebra. If each X_n is a compact metrizable space, then $\prod_{n=0}^{\infty} X_n$ is endowed with the product topology; in this case, $\prod_{n=0}^{\infty} X_n$ is compact

and metrizable, and its Borel σ-algebra coincides with the product of Borel σ-algebras on X_n. For a measurable space X, we denote by $\Delta(X)$ the set of probability measures over X, endowed with the σ-algebra generated by $\{\mu \in \Delta(X) \mid \mu(E) \geq p\}$ for each measurable subset $E \subseteq X$ and each $p \in [0,1]$. If X is a compact metric space, then $\Delta(X)$ is endowed with the weak-* topology; in this case, $\Delta(X)$ is compact and metrizable, and its Borel σ-algebra coincides with the σ-algebra on $\Delta(X)$ generated from the Borel σ-algebra on X. For notational simplicity, we sometimes write $\mu(x)$ for $\mu(\{x\})$.

We let I be a non-empty finite set of agents, and Z be a finite set of outcomes with $|Z| \geq 2$. To keep our language as standard as possible, we find it convenient to identify expected utility preferences with representations of those preferences in $\mathbb{R}^Z$. Thus, we say that for lotteries $p, p' \in \Delta(Z)$, an agent with "preference" $u_i \in \mathbb{R}^Z$ prefers p to p' if and only if

$$\sum_{z \in Z} p(z)\, u_i(z) \geq \sum_{z \in Z} p'(z)\, u_i(z).$$

3.2. Restrictions on conditional preferences

We will require that each agent i's conditional preferences over lotteries, given the opponents' type profiles, can be represented by a von Neumann–Morgenstern utility index within a given set $U_i^\bullet \subset \mathbb{R}^Z$ that satisfies

1. uniform ranking: there exists a pair of outcomes $\overline{z}, \underline{z} \in Z$ such that $u_i(\overline{z}) > u_i(\underline{z})$ for every $u_i \in U_i^\bullet$; we normalize each u_i so that $u_i(\overline{z}) = 1$ and $u_i(\underline{z}) = 0$;[8]
2. bounded utility: there exists $B_i \geq 1$ such that $|u_i(z)| \leq B_i$ for every $u_i \in U_i^\bullet$ and $z \in Z$ (given the above normalization).

Given $U_i^\bullet$ that satisfies the uniform ranking and bounded utility assumptions, we can embed $U_i^\bullet$ into a simplex $\mathrm{co}(U_i)$ with vertices (i.e., extreme points) $U_i = \{u_i^1, u_i^2, \ldots, u_i^{K_i}\}$ that satisfy the following properties:

1. unique representation: no two distinct utility indices in $\mathrm{co}(U_i)$ represent the same preference;
2. non-constant utility: no utility index in $\mathrm{co}(U_i)$ is constant;
3. linear independence: $u_i^2 - u_i^1, u_i^3 - u_i^1, \ldots, u_i^{K_i} - u_i^1$ are linearly independent.

Property 1 comes from the normalization among representations. Property 2 rules out complete indifference from $\mathrm{co}(U_i)$ and follows from uniform ranking. Property 3 is the linear independence assumption; it requires that every preference in $\mathrm{co}(U_i)$ can be uniquely represented as a convex combination of the extreme points. Our results hold for any $(U_i)_{i \in I}$ satisfying properties 1 through 3, and the uniform ranking and bounded utility assumptions are sufficient conditions for them to hold.

We can illustrate the embedding and why Property 3 holds by construction of extreme points of a simplex. If $|Z| \geq 3$ and agent i has uniform ranking between the first and second outcomes, choose $U_i = \{u_i^1, u_i^2, \ldots, u_i^{|Z|-1}\}$ such that

[8] For notational convenience, we require uniform ranking of a pair of pure outcomes. Our analysis would be unchanged even if we required uniform ranking of a pair of lotteries.

$$u_i^1 = (1, 0, -B_i, -B_i, \ldots, -B_i),$$
$$u_i^2 = (1, 0, C_i, -B_i, \ldots, -B_i),$$
$$u_i^3 = (1, 0, -B_i, C_i, \ldots, -B_i),$$
$$\vdots$$
$$u_i^{|Z|-1} = (1, 0, -B_i, -B_i, \ldots, C_i)$$

with sufficiently large C_i. (For example, let $C_i = (2|Z| - 5)B_i$.)

To illustrate the uniform ranking and bounded utility assumptions, we will describe how the conditional altruism example discussed in Section 2 fits into the framework of this Section. Suppose that there are two agents and three outcomes, $Z = \{\varnothing, 1, 2\}$, where the outcomes correspond to, respectively, no one getting the prize, agent 1 getting the prize, and agent 2 getting the prize. The set U_1 consists of two vectors $(0, 1, 0)$ and $(0, 1, 1)$ corresponding to, respectively, the extreme preferences where the agent 1 is indifferent between the other agent getting the prize (outcome 2) and no one getting the prize (outcome $\varnothing$) and where the agent is indifferent between the other agent getting the prize (outcome 2) and getting the prize herself (outcome 1). Symmetrically, U_2 consists of two vectors $(0, 0, 1)$ and $(0, 1, 1)$.

We can also use this example to illustrate how the set of allowable preferences can be generalized. For example, we could replace the two extreme preferences of agent 1 by $(0, 1, B_1)$ and $(0, 1, -B_1)$, for some large $B_1 \geq 1$. This allows for the possibility that agent 1 strictly prefers agent 2 getting the prize to getting the prize himself. And it allows a "spiteful" agent 1 who strictly prefers no one getting the prize to agent 2 getting the prize. This continues to satisfy the uniform ranking and bounded utility assumptions.

We will hold $(U_i)_{i \in I}$ fixed throughout our analysis, except in Sections 5.3–5.5. In Section 5.3, we will discuss a sense in which our main results are independent of the choice of $(U_i)_{i \in I}$.[9] Recall that while each $\text{co}(U_i)$ represents a set of expected utility preferences over lotteries, $\text{co}(U_i)$ is isomorphic to the set of probability distributions over its extreme points, $\Delta(U_i)$, and this will play an important role in our presentation. In particular, a preference can then conveniently be thought of as a probability distribution over "private states" U_i.

3.3. Type spaces

We first describe implicit, self-referential, type spaces allowing interdependent preferences. Given our embedding of each agent's utility indices within a simplex, it is convenient to represent his type as a probability distribution over private states (i.e., extreme points of that agent's possible preferences) and the opponents' type profiles.

A *type space based on* $(U_i)_{i \in I}$, $\mathcal{T} = (T_i, \mu_i)_{i \in I}$, consists of non-empty measurable spaces T_i of agent i's possible types, $T_{-i} \equiv \prod_{j \neq i} T_j$, and measurable mappings:

$$\mu_i : T_i \to \Delta(U_i \times T_{-i}),$$

where μ_i assigns a belief $\mu_i(t_i) \in \Delta(U_i \times T_{-i})$ to each type $t_i \in T_i$. We interpret the marginal probability distribution $\text{mrg}_{T_{-i}} \mu_i(t_i) \in \Delta(T_{-i})$ as type t_i's belief over the opponents' type profiles, and (a version of) the conditional probability distribution given the opponents' type profiles

[9] In Section 5.4, we will also discuss how to relax the uniform ranking and bounded utility assumptions.

$\mu_i(t_i)(\cdot \mid \cdot)\colon T_{-i} \to \Delta(U_i)$ as utility indices that represent type t_i's conditional preferences given the opponents' type profiles. Thus $\mu_i(t_i)$ represents type t_i's preferences over Anscombe–Aumann acts defined on others' types. Note that correlation in $\mu_i(t_i) \in \Delta(U_i \times T_{-i})$ is essential, as it allows us to express the dependency of type t_i's preferences on the opponents' type profiles. Also note that unlike the "interdependent payoff universal belief space" T_{IP} in Section 2, our type space does not specify type t_i's conditional preferences given the opponents' type profiles that he attaches probability zero to because the conditional probability distribution $\mu_i(t_i)(\cdot \mid \cdot)$ is identified only *almost surely* with respect to $\text{mrg}_{T_{-i}}\mu_i(t_i)$.

3.4. The universal type space

Because $\text{co}(U_i)$ is isomorphic to $\Delta(U_i)$, we can interpret U_i as a finite set of extreme "payoff states" and $u_i \in \text{co}(U_i)$ as a probability distribution over those payoff states. Thus, we can treat $\mathcal{T} = (T_i, \mu_i)_{i\in I}$ formally as a belief type space, where agents have beliefs and higher order beliefs over private state spaces $(U_i)_{i\in I}$. With minor modifications of Mertens and Zamir (1985) and Brandenburger and Dekel (1993), both of which use a common state space, we define the universal type space $\mathcal{T}^* = (T_i^*, \mu_i^*)_{i\in I}$ based on $(U_i)_{i\in I}$, where T_i^* is the set of all belief hierarchies of agent i that satisfy coherence and common certainty of coherence, which is nonempty, compact, and metrizable, and μ_i^* is the natural homeomorphism

$$\mu_i^*\colon T_i^* \to \Delta(U_i \times T_{-i}^*).$$

Note that the belief hierarchy of each type in the universal type space coincides with the type itself. Moreover, for every type space $\mathcal{T} = (T_i, \mu_i)_{i\in I}$ based on $(U_i)_{i\in I}$, the mapping that maps each type in T_i to his hierarchy of beliefs:

$$\hat{\mu}_i\colon T_i \to T_i^*$$

preserves the belief structure, i.e.,

$$\mu_i^*(\hat{\mu}_i(t_i))(E) = \mu_i(t_i)(\{(u_i, t_{-i}) \in U_i \times T_{-i} \mid (u_i, (\hat{\mu}_j(t_j))_{j\neq i}) \in E\})$$

for every measurable subset $E \subseteq U_i \times T_{-i}^*$. We sometimes write $\hat{\mu}_i(\cdot; \mathcal{T})$ to emphasize the underlying type space. We will refer to T_i^* as the set of interdependent preference hierarchies, to highlight the interpretation of this mathematical object in this paper. In particular, for each type t_i, belief hierarchy $\hat{\mu}_i(t_i)$ represents his interdependent preference hierarchy in such a way that the first order belief in $\hat{\mu}_i(t_i)$ represents his preference over lotteries, the second order belief in $\hat{\mu}_i(t_i)$ represents his preference over Anscombe–Aumann acts defined over profiles of other agents' first order beliefs (which represent their preferences over lotteries), etc....

3.5. Interim correlated rationalizability

A *mechanism* (or *game form*) is given by $\mathcal{M} = ((M_i)_{i\in I}, O)$, where M_i is a non-empty set of messages (actions) available to agent i, $M = \prod_{i\in I} M_i$, and $O\colon M \to \Delta(Z)$ is the outcome function. In this mechanism, agents send messages $m = (m_i)_{i\in I} \in M$ simultaneously, and the mechanism assigns an outcome z with probability $O(m)(z)$. A mechanism $\mathcal{M} = ((M_i)_{i\in I}, O)$ is *finite* if M_i is finite for every $i \in I$. Except in Section 4.2 and Appendix C.3, where we will formulate technical lemmas in terms of single-agent infinite mechanisms, we restrict ourselves to finite mechanisms.

A type space $\mathcal{T} = (T_i, \mu_i)_{i \in I}$ and a mechanism $\mathcal{M} = ((M_i)_{i \in I}, O)$ together define a Bayesian game. We will later define and discuss equilibrium and other solution concepts for this game. However, it is useful to first discuss a version of interim correlated rationalizability (ICR) in Dekel et al. (2007) adapted to the present setting. We differ from DFM with respect to the structure of Bayesian games: we have private state spaces $(U_i)_{i \in I}$ while DFM have a common state space. We also differ with respect to the interpretation: here, states represent extreme points of utility indices that represent each agent's possible conditional preferences while at least in a leading interpretation of DFM, states represent external events on which the payoffs of the game are conditioned.[10] Formally, given a type space $\mathcal{T}$ and a finite mechanism $\mathcal{M}$, ICR is defined by induction as follows. The induction is initialized with

$$R_i^0(t_i) = M_i,$$

with the inductive step defined by:

$$R_i^{n+1}(t_i) = \left\{ m_i \in M_i \;\middle|\; \begin{array}{l} \text{there exists } \nu_i \in \Delta(M_{-i} \times U_i \times T_{-i}) \text{ s.t.} \\ \text{(i) } \nu_i(\{(m_{-i}, u_i, t_{-i}) \in M_{-i} \times U_i \times T_{-i} \mid m_j \in R_j^n(t_j) \\ \qquad \text{for every } j \neq i\}) = 1, \\ \text{(ii) } \operatorname{mrg}_{U_i \times T_{-i}} \nu_i = \mu_i(t_i), \\ \text{(iii) } \int_{M_{-i} \times U_i \times T_{-i}} \sum_{z \in Z} u_i(z)(O(m_i, m_{-i})(z) - O(m_i', m_{-i})(z)) \\ \qquad \times \nu_i(dm_{-i}, du_i, dt_{-i}) \geq 0 \text{ for every } m_i' \in M_i \end{array} \right\},$$

and the limit is defined by:

$$R_i(t_i) = \bigcap_{n=0}^{\infty} R_i^n(t_i).$$

Note that the inductive step is well defined since we can show inductively that $\{(m_i, t_i) \in M_i \times T_i \mid m_i \in R_i^n(t_i)\}$ is measurable in $M_i \times T_i$ for every $i \in I$ and $n \geq 0$. We say that m_i is *interim correlated rationalizable for* t_i if $m_i \in R_i(t_i)$. We sometimes write $R_i(t_i; \mathcal{T}, \mathcal{M})$ to emphasize the underlying type space and mechanism.

As in Dekel et al. (2007, Proposition 1 and Corollary 2), ICR depends only on hierarchies of interdependent preferences.

Proposition 1. *For every type space $\mathcal{T} = (T_i, \mu_i)_{i \in I}$ based on $(U_i)_{i \in I}$, every agent $i \in I$, and every type $t_i \in T_i$, we have*

$$R_i(t_i; \mathcal{T}, \mathcal{M}) = R_i(\hat{\mu}_i(t_i); \mathcal{T}^*, \mathcal{M})$$

for every finite mechanism $\mathcal{M} = ((M_i)_{i \in I}, O)$.

We omit the proof of this Proposition, which requires a cosmetic modification of DFM's proofs to incorporate private state spaces.

[10] This difference in interpretation will be important in Sections 5.1 and 5.2. Correlation in an agent's conjecture about that agent's private state and other agents' actions corresponds to interdependency of that agent's preferences on others agents' actions. In this sense, ICR can be seen as even more permissive in the present context. See Morris and Takahashi (2012) for more on the foundations and interpretations of these solution concepts.

4. Strategic distinguishability

4.1. Strategic distinguishability for ICR

For any interim solution concept, we say that two types are *strategically indistinguishable* if their sets of solutions have a non-empty intersection for every finite mechanism. In this terminology, the following Theorem establishes that hierarchies of interdependent preferences characterize strategic distinguishability for ICR.

Theorem 1. *For every pair of type spaces $\mathcal{T} = (T_i, \mu_i)_{i\in I}$ and $\mathcal{T}' = (T_i', \mu_i')_{i\in I}$ based on $(U_i)_{i\in I}$, every agent $i \in I$, and every pair of types $t_i \in T_i$ and $t_i' \in T_i'$, the following two conditions are equivalent:*

1. $\hat{\mu}_i(t_i; \mathcal{T}) = \hat{\mu}_i(t_i'; \mathcal{T}')$;
2. $R_i(t_i; \mathcal{T}, \mathcal{M}) \cap R_i(t_i'; \mathcal{T}', \mathcal{M}) \neq \emptyset$ *for every finite mechanism* $\mathcal{M}$.

Note that $1 \Rightarrow 2$ follows from Proposition 1 and the nonemptyness of ICR. $2 \Rightarrow 1$ follows from the following Proposition, which we will show by establishing the contrapositive $\neg 1 \Rightarrow \neg 2$. Let d_i^* be a metric compatible with the product topology on T_i^*.

Proposition 2. *For every $\varepsilon > 0$, there exists a finite mechanism $\mathcal{M}$ such that*

$$d_i^*(\hat{\mu}_i(t_i; \mathcal{T}), \hat{\mu}_i(t_i'; \mathcal{T}')) > \varepsilon \Rightarrow R_i(t_i; \mathcal{T}, \mathcal{M}) \cap R_i(t_i'; \mathcal{T}', \mathcal{M}) = \emptyset$$

for every pair of type spaces $\mathcal{T} = (T_i, \mu_i)_{i\in I}$ and $\mathcal{T}' = (T_i', \mu_i')_{i\in I}$ based on $(U_i)_{i\in I}$, every agent $i \in I$, and every pair of types $t_i \in T_i$ and $t_i' \in T_i'$.

The Proposition proves a little more than what is needed to prove Theorem 1: it shows that if we fix a metric d_i^* and $\varepsilon > 0$, we can strategically distinguish all types that are at least ε apart using the *same mechanism*. In the remainder of this Subsection, we describe the mechanism used to prove this result, which is the main technical contribution of the paper.

The strategy of proof is as follows. If two types are ε apart in the metric compatible with the product topology on T_i^*, then there must exist $\overline{\varepsilon} > 0$ and N such that the types' Nth order preferences are at least $\overline{\varepsilon}$ apart. We will choose "accuracy" levels $0 < \varepsilon_0 \leq \varepsilon_1 \leq \cdots \leq \varepsilon_N$. For each agent i and $n \geq 1$, agent i will report an element of an ε_{n-1}-dense finite subset of his possible nth order preferences. For each agent i and $n \geq 1$, there will be a component of the mechanism, chosen with positive probability, that will pick an outcome as a function of agent i's report about his nth order preference and the other agents' reports about their $(n-1)$th and lower order preferences. The mechanism will have the property that as long as the other agents' reports are within ε_{n-1} of their true preferences, then agent i's best responses are within ε_n of his true nth order preference. Using this property inductively, we will show that each agent's ICR reports about his nth order preference are within ε_n of his true nth order preference.

The last step of the argument uses a robust scoring rule described in the next Subsection. We show that, for every $\varepsilon > 0$, we can find $\delta > 0$ and a scoring rule that gives the agent an incentive to report preferences within ε of his true preference even if the outcomes of the scoring rule may

be arbitrarily perturbed within δ. This Lemma can then be iteratively applied to construct the mechanism used in the main proof.

Abreu and Matsushima (1992) and Dekel et al. (2006) follow similar arguments up until the second last step. AM exploit the finiteness of the type space. They can choose an $\varepsilon > 0$ such that the $(j, n+1)$th component occurs with probability at most ε times that of the (i, n)th component. Now $\varepsilon > 0$ can be chosen uniformly small enough so that agents can be strictly incentivized to report their true preferences exactly at every order.[11] DFM allow for arbitrary, possibly infinite, type spaces, so it is not possible to find a uniform ε that makes the argument in AM work. In DFM, it is necessary to have each agent report an element of a finite grid of beliefs at every order. But payoffs can be chosen independently across agents, so it is possible to do the approximation inductively. Because neither proof strategy is available in our setting, we need a novel robust scoring rule to make the argument work.

4.2. The robust scoring rule

As a preliminary step, we first analyze a single-agent mechanism that reveals his state-dependent preferences. In this Subsection, fix a compact metric space X of states with metric d. Let d_Δ be a metric compatible with the weak-* topology over $\Delta(U \times X)$. Let $F(X)$ be the set of (Anscombe–Aumann) acts over X, i.e., the set of measurable functions $f : X \to \Delta(Z)$. Then each $\mu \in \Delta(U \times X)$ uniquely represents a state-dependent preference over $F(X)$. That is, the agent with preference μ weakly prefers f to f' if and only if

$$\int_{U \times X} \sum_{z \in Z} u(z)(f(x)(z) - f'(x)(z))\mu(du, dx) \geq 0.$$

We define the choice function with respect to μ:

$$C_\mu(f, f') = \begin{cases} f & \text{if } \mu \text{ weakly prefers } f \text{ to } f', \\ f' & \text{if } \mu \text{ strictly prefers } f' \text{ to } f, \end{cases}$$

for every $f, f' \in F(X)$.

Let $F_c(X) \subseteq F(X)$ be the set of continuous acts over X. Since X is a compact metric space, by the Stone–Weierstrass theorem, there exists a countable dense subset $F = \{f_1, f_2, \ldots\} \subset F_c(X)$ in the sup norm. Fix such an F.

We consider the following direct mechanism $\mathcal{M}^0 = (M^0, O^0)$ for a single agent with message set $M^0 = \Delta(U \times X)$ and outcome function $O^0 : M^0 \times X \to \Delta(Z)$ given by

$$O^0(m, x)(z) = \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} 2^{-k-l} C_m(f_k, f_l)(x)(z), \tag{1}$$

for each realized state $x \in X$ and each reported preference $m \in M^0$. Under the mechanism $\mathcal{M}^0$, the agent reports a preference. Then the mechanism randomly draws an ordered pair (f_k, f_l) of acts from F with probability 2^{-k-l}, and then assigns the agent with the preferred act according to the reported preference.[12]

[11] In the related work of Bergemann and Morris (2009b), there is a finite set of possible "payoff types" and an analogous trick can be applied.

[12] Note that $\mathcal{M}^0$ is not a finite mechanism. The mechanism we will construct in the next Subsection to prove Proposition 2, however, is finite.

In Lemma 1 below, we show that truth telling is optimal in $\mathcal{M}^0$ for every preference. Indeed, by invoking the compactness of X, we show a "robust" version of optimality: in every mechanism close to $\mathcal{M}^0$, the agent strictly prefers reporting his approximately true preferences to reporting any other.

Recall that for each message m, $O^0(m,\cdot)$ is an act over X, which determines an outcome z with probability $O^0(m,x)(z)$ when nature chooses $x \in X$. We consider two sources of perturbations to this act. First, with small probability, the outcome may not be chosen according to $O^0(m,x)$. Formally, for each $\delta > 0$ and measurable space Ω, we consider a perturbed outcome function $O\colon M^0 \times X \times \Omega \to \Delta(Z)$ such that

$$\|O(\cdot,\cdot,\omega) - O^0\| \equiv \sup_{m \in M^0, x \in X, z \in Z} |O(m,x,\omega)(z) - O^0(m,x)(z)| \leq \delta$$

for every $\omega \in \Omega$. Second, when nature is supposed to choose x, nature may instead choose x' in a neighborhood of x. Formally, for each $\delta > 0$, $\mu \in \Delta(U \times X)$, and measurable space Ω, let

$$\Delta_{\delta,\mu}(U \times X \times \Omega) = \left\{ \mu'' \in \Delta(U \times X \times \Omega) \;\middle|\; \begin{array}{l} \text{there exists } \mu' \in \Delta(U \times X \times X' \times \Omega) \\ \quad \text{with } X' = X \text{ s.t.} \\ \text{(i) } \mu'(U \times \{(x,x') \mid d(x,x') \leq \delta\} \times \Omega) = 1, \\ \text{(ii) } \mathrm{mrg}_{U \times X} \mu' = \mu, \\ \text{(iii) } \mathrm{mrg}_{U \times X' \times \Omega} \mu' = \mu'' \end{array} \right\}, \tag{2}$$

be the set of preferences over noisy acts induced by the original preference μ.[13]

Lemma 1. *For every $\varepsilon > 0$, there exists $\delta > 0$ such that the following is true for every preference $\mu \in \Delta(U \times X)$, every pair of messages m, m', every measurable space Ω, and every perturbed outcome function $O\colon M^0 \times X \times \Omega \to \Delta(Z)$: if $d_\Delta(\mu, m) \leq \delta$, $d_\Delta(\mu, m') > \varepsilon$, and $\|O(\cdot,\cdot,\omega) - O^0\| \leq \delta$ for every $\omega \in \Omega$, then every preference in $\Delta_{\delta,\mu}(U \times X \times \Omega)$ strictly prefers $O(m,\cdot,\cdot)$ to $O(m',\cdot,\cdot)$.*

The proof is in Appendix A.

We call the mechanism $\mathcal{M}^0 = (M^0, O^0)$ a "robust scoring rule" because it elicits the agent's preference in the following robust way. For each fixed preference of the agent, every message that is close to that fixed preference is strictly preferred by the agent to every message which is at some distance from that preference, under every mechanism close to our scoring rule and every preference close to the fixed preference. The arguments of AM and DFM also use scoring rules, and we extend their use to elicit hierarchies of interdependent preferences or beliefs. For their arguments, it is enough to use a standard scoring rule. The extra generality of our setting necessitates the use of a robust scoring rule.

4.3. Proof of Proposition 2

We first prepare for notations for belief hierarchies. Recall that we follow the standard procedure and construct the universal type space $\mathcal{T}^* = (T_i^*, \mu_i^*)_{i \in I}$ of belief hierarchies. Specifically, for each $i \in I$, letting $H_{i,0} = \{*\}$ be initialized with a single element, we denote by

[13] We introduce X' as a copy of X to notationally distinguish the marginal of μ' on X (the "first X") and on X' (the "second X").

$H_{i,n} = H_{i,n-1} \times \Delta(U_i \times H_{-i,n-1}) = \prod_{k=0}^{n-1} \Delta(U_i \times H_{-i,k})$ the set of higher order beliefs up to nth order for each $n \geq 1$. Then we can construct the universal type space $T_i^* \subset \prod_{n=0}^{\infty} \Delta(U_i \times H_{-i,n})$ as the set of agent i's belief hierarchies satisfying coherence, in the sense that lower order beliefs are marginals of higher order beliefs, and common certainty of coherence. Recall that d_i^* is a metric compatible with the product topology on T_i^*. Let $d_{i,n}$ be a metric compatible with the topology on the set of agent i's nth order beliefs, $\Delta(U_i \times H_{-i,n-1})$.

Fix any $\varepsilon > 0$. By the definition of the product topology, there exist $\bar{\varepsilon} > 0$ and $N \in \mathbb{N}$ such that, for every $(t_{i,n})_{n=1}^{\infty}, (t'_{i,n})_{n=1}^{\infty} \in T_i^*$, if $d_i^*((t_{i,n})_{n=1}^{\infty}, (t'_{i,n})_{n=1}^{\infty}) > \varepsilon$, then there exists some $n \leq N$ such that $d_{i,n}(t_{i,n}, t'_{i,n}) > \bar{\varepsilon}$. Pick such $\bar{\varepsilon}$ and N.

For each $i \in I$ and $n \leq N$, we apply Lemma 1 by substituting

$$X = H_{-i,n-1} = \prod_{j \neq i} \prod_{k=0}^{n-2} \Delta(U_j \times H_{-j,k}),$$

$$d = \max_{j \neq i, 1 \leq k \leq n-1} d_{j,k},$$

$$d_{\Delta} = d_{i,n}.$$

Pick a countable dense subset of $F_c(H_{-i,n-1})$, and define $O_{i,n}^0 \colon \Delta(U_i \times H_{-i,n-1}) \times H_{-i,n-1} \to \Delta(Z)$ as in (1). By Lemma 1, there exist $0 < \varepsilon_0 \leq \varepsilon_1 \leq \cdots \leq \varepsilon_N \leq \bar{\varepsilon}/2$ such that if $d_{i,n}(t_{i,n}, m_{i,n}) \leq \varepsilon_{n-1}$, $d_{i,n}(t_{i,n}, m'_{i,n}) > \varepsilon_n$, and $\| O_{i,n}(\cdot, \cdot, \omega) - O_{i,n}^0 \| \leq \varepsilon_{n-1}$ for every $\omega \in \Omega$, then every preference in $\Delta_{\varepsilon_{n-1}, t_{i,n}}(U_i \times H_{-i,n-1} \times \Omega)$ strictly prefers $O_{i,n}(m_{i,n}, \cdot, \cdot)$ to $O_{i,n}(m'_{i,n}, \cdot, \cdot)$.

We define a finite mechanism $\mathcal{M} = ((M_i)_{i \in I}, O)$ as follows. For each $i \in I$ and $n \leq N$, let $M_{i,n}$ be any ε_{n-1}-dense finite subset of $\Delta(U_i \times H_{-i,n-1})$ with respect to $d_{i,n}$, and $M_i = \prod_{n=1}^{N} M_{i,n}$. Define $O \colon M \to \Delta(Z)$ by

$$O(m)(z) = \frac{1-\delta}{|I|(1-\delta^N)} \sum_{i \in I} \sum_{n=1}^{N} \delta^{n-1} O_{i,n}^0(m_{i,n}, m_{-i,1}, \ldots, m_{-i,n-1})(z)$$

for each $m \in M$ and $z \in Z$, where $\delta > 0$ is small enough to satisfy $(1-\delta)/\delta \geq (|I|-1)(1-\varepsilon_0)/\varepsilon_0$.

Lemma 2. *For every type space $\mathcal{T} = (T_i, \mu_i)_{i \in I}$ based on $(U_i)_{i \in I}$, every agent $i \in I$, and every type $t_i \in T_i$, we have*

$$m_i \in R_i^n(t_i; \mathcal{T}, \mathcal{M}) \Rightarrow d_{i,n}(\hat{\mu}_{i,n}(t_i), m_{i,n}) \leq \varepsilon_n$$

for every $n \leq N$.

The proof of this Lemma is in Appendix A. We can now complete the proof of Proposition 2.

Proof of Proposition 2. Let $\mathcal{M}$ be the finite mechanism defined above. Pick any pair of type spaces $\mathcal{T}$ and $\mathcal{T}'$ based on $(U_i)_{i \in I}$, $i \in I$, $t_i \in T_i$, and $t'_i \in T'_i$. Suppose that there exists $m_i = (m_{i,1}, \ldots, m_{i,N}) \in R_i(t_i; \mathcal{T}, \mathcal{M}) \cap R_i(t'_i; \mathcal{T}', \mathcal{M})$. For every $n \leq N$, since $a_i \in R_i^n(t_i; \mathcal{T}, \mathcal{M}) \cap R_i^n(t'_i; \mathcal{T}', \mathcal{M})$, we have

$$d_{i,n}(\hat{\mu}_{i,n}(t_i; \mathcal{T}), \hat{\mu}_{i,n}(t'_i; \mathcal{T}')) \leq d_{i,n}(\hat{\mu}_{i,n}(t_i; \mathcal{T}), m_{i,n}) + d_{i,n}(\hat{\mu}_{i,n}(t'_i; \mathcal{T}'), m_{i,n}) \leq 2\varepsilon_n \leq \bar{\varepsilon}$$

by Lemma 2. Thus, $d_i^*(\hat{\mu}_i(t_i; \mathcal{T}), \hat{\mu}_i(t'_i; \mathcal{T}')) \leq \varepsilon$. □

4.4. Strategic distinguishability for equilibrium

Our analysis thus far concerned the solution concept of ICR. We now change our focus to equilibrium. Given a type space $\mathcal{T}$ and a finite mechanism $\mathcal{M}$, we say that a profile $\sigma = (\sigma_i)_{i\in I}$ of measurable behavioral strategies $\sigma_i : T_i \to \Delta(M_i)$ is an *equilibrium* if

$$\int_{U_i\times T_{-i}} \sum_{m_{-i}\in M_{-i}} \sum_{z\in Z} u_i(z)(O(m_i, m_{-i})(z) - O(m_i', m_{-i})(z)) \times \left(\prod_{j\neq i} \sigma_j(t_j)(m_j)\right) \mu_i(t_i)(du_i, dt_{-i}) \geq 0$$

for every $i \in I$, every $t_i \in T_i$ and every $m_i, m_i' \in M_i$ with $\sigma_i(t_i)(m_i) > 0$. We denote by $E_i(t_i; \mathcal{T}, \mathcal{M})$ the set of actions played by type t_i with positive probability in some equilibrium.

We have the following (called the "pull-back property" in Friedenberg and Meier (2015)):

Proposition 3. *For every type space $\mathcal{T} = (T_i, \mu_i)_{i\in I}$ based on $(U_i)_{i\in I}$, every agent $i \in I$, and every type $t_i \in T_i$, we have*

$$E_i(t_i; \mathcal{T}, \mathcal{M}) \supseteq E_i(\hat{\mu}_i(t_i); \widehat{\mathcal{T}}, \mathcal{M})$$

for every finite mechanism $\mathcal{M} = ((M_i)_{i\in I}, O)$, where $\widehat{\mathcal{T}} = (\widehat{T}_i, \mu_i^*|_{\widehat{T}_i})_{i\in I}$ is a belief closed subspace of the universal type space $\mathcal{T}^* = (T_i^*, \mu_i^*)_{i\in I}$ with $\widehat{T}_i = \hat{\mu}_i(T_i)$ for each $i \in I$.

The proof is in Appendix A.

We say that a type space $\mathcal{T} = (T_i, \mu_i)_{i\in I}$ is *finite* (or *countable*) if T_i is finite (or countable) for every $i \in I$. A type is a *finite* (or a *countable*) type if it lies in a finite (or countable) type space. Equilibria do not always exist on uncountable type spaces: see Simon (2003), Friedenberg and Meier (2015) and Hellman (2014). However, since the mechanism is finite, the existence of equilibria is guaranteed on any countable type space.[14] This gives us:

Theorem 2. *For every pair of countable type spaces $\mathcal{T} = (T_i, \mu_i)_{i\in I}$ and $\mathcal{T}' = (T_i', \mu_i')_{i\in I}$ based on $(U_i)_{i\in I}$, every agent $i \in I$, and every pair of types $t_i \in T_i$ and $t_i' \in T_i'$, the following two conditions are equivalent:*

1. $\hat{\mu}_i(t_i; \mathcal{T}) = \hat{\mu}_i(t_i'; \mathcal{T}')$;
2. $E_i(t_i; \mathcal{T}, \mathcal{M}) \cap E_i(t_i'; \mathcal{T}', \mathcal{M}) \neq \emptyset$ *for every finite mechanism* $\mathcal{M}$.

[14] To see this, let $T_i \subseteq \mathbb{N}$ without loss of generality. Then the set $(\Delta(M_i))^{T_i}$ of behavior strategies of agent i is a nonempty, compact and convex subset of a locally convex and Hausdorff topological vector space $\mathbb{R}^{T_i\times M_i}$ (endowed with the product topology), and agent i's payoff function

$$v_i(\sigma) = \sum_{t_i\in T_i} 2^{-t_i} \sum_{u_i\in U_i, t_{-i}\in T_{-i}} \sum_{m\in M} \sum_{z\in Z} u_i(z) O(m)(z) \left(\prod_{j\in I} \sigma_j(t_j)(m_j)\right) \mu_i(t_i)(u_i, t_{-i})$$

is affine in σ_i and continuous in σ (under the product topology) by the Lebesgue convergence theorem. Thus, the existence of equilibria follows from Berge's maximum theorem and the Kakutani–Fan–Glicksberg fixed-point theorem in the usual way.

Proof. For the $1 \Rightarrow 2$ direction, we denote $t_i^* := \hat{\mu}_i(t_i; \mathcal{T}) = \hat{\mu}_i(t_i'; \mathcal{T}')$. By Proposition 3, we have

$$E_i(t_i; \mathcal{T}, \mathcal{M}) \cap E_i(t_i'; \mathcal{T}', \mathcal{M}) \supseteq E_i(t_i^*; \widehat{\mathcal{T}}, \mathcal{M}) \cap E_i(t_i^*; \widehat{\mathcal{T}}', \mathcal{M}),$$

where $\widehat{\mathcal{T}} = (\widehat{T}_i, \mu_i^*|_{\widehat{T}_i})_{i \in I}$ and $\widehat{\mathcal{T}}' = (\widehat{T}_i', \mu_i^*|_{\widehat{T}_i'})_{i \in I}$ are countable belief closed subspaces of the universal type space $\mathcal{T}^* = (T_i^*, \mu_i^*)_{i \in I}$ with $\widehat{T}_i = \widehat{\mu}_i(T_i; \mathcal{T})$ and $\widehat{T}_i' = \widehat{\mu}_i(T_i'; \mathcal{T}')$ for each $i \in I$. Let $\widehat{\mathcal{T}}'' = (\widehat{T}_i \cap \widehat{T}_i', \mu_i^*|_{\widehat{T}_i \cap \widehat{T}_i'})_{i \in I}$. Applying Proposition 3 to inclusion maps from $\widehat{\mathcal{T}}''$ to $\widehat{\mathcal{T}}$ and to $\widehat{\mathcal{T}}'$, we can show that the restriction of every equilibrium of $(\widehat{\mathcal{T}}, \mathcal{M})$ or of $(\widehat{\mathcal{T}}', \mathcal{M})$ to $\widehat{\mathcal{T}}''$ is also an equilibrium of $(\widehat{\mathcal{T}}'', \mathcal{M})$. Conversely, by the fixed-point argument sketched in footnote 14, we can show that every equilibrium of $(\widehat{\mathcal{T}}'', \mathcal{M})$ extends to equilibria of $(\widehat{\mathcal{T}}, \mathcal{M})$ and of $(\widehat{\mathcal{T}}', \mathcal{M})$ (called the "extension property" in Friedenberg and Meier (2015)). Thus, both $E_i(t_i^*; \widehat{\mathcal{T}}, \mathcal{M})$ and $E_i(t_i^*; \widehat{\mathcal{T}}', \mathcal{M})$ are equal to $E_i(t_i^*; \widehat{\mathcal{T}}'', \mathcal{M})$, which is nonempty.

The $2 \Rightarrow 1$ direction, or its contrapositive $\neg 1 \Rightarrow \neg 2$, follows from Proposition 2 and the fact that equilibrium is a refinement of ICR. □

The $1 \Rightarrow 2$ direction of Theorem 2 has been known in the literature. For example, in the setting with common certainty of conditional preferences over lotteries, Yildiz (2015) shows the existence of an invariant equilibrium defined over all finite types, which depends only on Mertens–Zamir belief hierarchies about external states.[15]

It is immediate from Theorems 1 and 2 that hierarchies of interdependent preferences characterize strategic distinguishability for other interim solution concepts that are coarser than equilibrium and finer than ICR, for example, interim independent rationalizability.

5. Extensions

5.1. Incorporating external states

We have so far considered "uncontingent" mechanisms $\mathcal{M} = ((M_i)_{i \in I}, O)$ with $O: M \to \Delta(Z)$, where agents' messages alone determine outcomes. We modeled agents' interdependent preferences, which entailed modeling the agents' incomplete information about each others' preferences. We showed that strategic distinguishability – using uncontingent mechanisms – was characterized by hierarchies of interdependent preferences.

However, game theorists often talk about incomplete information about external states, which we shall denote by $\theta \in \Theta$ (instead of or in addition to "private states" that we have introduced to express interdependent preferences). For simplicity, we assume Θ to be finite. Obviously, it will not be possible to elicit agents' beliefs and higher order beliefs about external states without allowing for richer mechanisms that assign outcomes contingent on those external states. Thus we consider "Θ-contingent" mechanisms $\mathcal{M} = ((M_i)_{i \in I}, O)$ with $O: M \times \Theta \to \Delta(Z)$, where the domain of the outcome function is extended to $M \times \Theta$. With this richer class of mechanisms, we will be able to achieve a finer strategic distinction of types, since these external states may also impact preferences, and beliefs and higher order beliefs about them may also be revealed. We excluded discussion of such external states earlier because they were incidental to our primary exercise of characterizing strategic distinguishability for interdependent preferences. But reporting this extension now allows us to connect our result to those of DFM, according to their

[15] See Section 5.2 for an interpretation of common certainty of conditional preferences in our setting.

original interpretation, in an exact way (see the next Subsection). The results and proofs do not change, once we alter our definitions of type spaces, mechanisms and solution concepts to reflect Θ in the appropriate way. Thus, we will merely state how the definitions must be changed in order for our previous results to hold as stated.

A type space $\mathcal{T} = (T_i, \mu_i)_{i\in I}$ now consists of non-empty measurable spaces T_i of agent i's possible types and measurable mappings $\mu_i : T_i \to \Delta(U_i \times \Theta \times T_{-i})$, i.e., a belief type space over private states and external states $(U_i \times \Theta)_{i\in I}$ representing agents' higher order preferences, their beliefs and higher order beliefs about Θ, and the interaction of the two. The universal type space based on $(U_i \times \Theta)_{i\in I}$, $\mathcal{T}^* = (T_i^*, \mu_i^*)_{i\in I}$, is constructed with the homeomorphism $\mu_i^* : T_i^* \to \Delta(U_i \times \Theta \times T_{-i}^*)$. For every type space $\mathcal{T} = (T_i, \mu_i)_{i\in I}$ based on $(U_i \times \Theta)_{i\in I}$, the mapping $\hat{\mu}_i : T_i \to T_i^*$ maps each type in T_i to its hierarchy of beliefs over $(U_i \times \Theta)_{i\in I}$. A Θ-*contingent mechanism* $\mathcal{M} = ((M_i)_{i\in I}, O)$ consists of non-empty sets M_i of messages available to agent i and the outcome function $O : M \times \Theta \to \Delta(Z)$. Given a type space $\mathcal{T} = (T_i, \mu_i)_{i\in I}$ and a finite Θ-contingent mechanism $\mathcal{M} = ((M_i)_{i\in I}, O)$, we define ICR by

$$R_i^0(t_i) = M_i,$$

$$R_i^{n+1}(t_i) = \left\{ m_i \in M_i \;\middle|\; \begin{array}{l} \text{there exists } \nu_i \in \Delta(M_{-i} \times U_i \times \Theta \times T_{-i}) \text{ s.t.} \\ \text{(i) } \nu_i(\{(m_{-i}, u_i, \theta, t_{-i}) \mid m_j \in R_j^n(t_j) \text{ for every } j \neq i\}) = 1, \\ \text{(ii) } \operatorname{mrg}_{U_i\times\Theta\times T_{-i}} \nu_i = \mu_i(t_i), \\ \text{(iii) } \int_{M_{-i}\times U_i\times\Theta\times T_{-i}} \sum_{z\in Z} u_i(z)(O(m_i, m_{-i}, \theta)(z) \\ \qquad - O(m_i', m_{-i}, \theta)(z))\nu_i(dm_{-i}, du_i, d\theta, dt_{-i}) \geq 0 \\ \qquad \text{for every } m_i' \in M_i \end{array} \right\},$$

$$R_i(t_i) = \bigcap_{n=0}^{\infty} R_i^n(t_i).$$

A profile $\sigma = (\sigma_i)_{i\in I}$ of measurable behavioral strategies $\sigma_i : T_i \to \Delta(M_i)$ is an equilibrium if

$$\int_{U_i\times\Theta\times T_{-i}} \sum_{m_{-i}\in M_{-i}} \sum_{z\in Z} u_i(z)(O(m_i, m_{-i}, \theta)(z) - O(m_i', m_{-i}, \theta)(z)) \times \left(\prod_{j\neq i} \sigma_j(t_j)(m_j)\right) \mu_i(t_i)(du_i, d\theta, dt_{-i}) \geq 0$$

for every $i \in I$, every $t_i \in T_i$, and every $m_i, m_i' \in M_i$ with $\sigma_i(t_i)(m_i) > 0$.

Now Theorems 1 and 2 remain true after replacing "based on $(U_i)_{i\in I}$" by "based on $(U_i \times \Theta)_{i\in I}$" and "mechanism" by "$\Theta$-contingent mechanism" and interpreting "$\hat{\mu}_i(t_i; \mathcal{T})$" and "$\hat{\mu}_i(t_i'; \mathcal{T}')$" as hierarchies of beliefs over $(U_i \times \Theta)_{i\in I}$. Our previous analysis corresponds to the special case where Θ is a singleton.

5.2. *Common certainty of conditional preferences*

We now maintain the extension incorporating external states (from the previous Subsection), but impose the restriction that there is "common certainty of conditional preferences," i.e., there is common certainty of how each outcome translates into a von Neumann–Morgenstern utility index. This corresponds to the setting of DFM, where it is implicitly assumed that there is com-

mon certainty of the payoffs associated with an action profile and an external state. This gives us one way of formally relating our results to those in DFM.[16]

We say that there is *common certainty of conditional preferences* if each U_i is a singleton $\{u_i\}$, where $u_i \in \mathbb{R}^Z$ is not constant over Z. Under common certainty of conditional preferences, there is uncertainty and higher order uncertainty about external states but no uncertainty about conditional preferences. Thus, the universal type space is simply the Mertens–Zamir universal type space, corresponding to the set of belief hierarchies about external states Θ satisfying coherence and common certainty of coherence.

Given that each U_i is a singleton, picking a contingent mechanism $\mathcal{M} = ((M_i)_{i\in I}, O)$ with $O\colon M \times \Theta \to \Delta(Z)$ is equivalent to picking a game with incomplete information about Θ (a specification of payoffs as a function of message/action profiles and external states), with the proviso that the set of feasible payoff vectors is given by the convex hull of the set of payoff vectors that can arise from some given outcome. Write V for the set of payoff profiles that can be induced by some lottery over outcomes, so that

$$V = \operatorname{conv}\{(u_i(z))_{i\in I} \in \mathbb{R}^I \mid z \in Z\}.$$

Now consider a game $\mathcal{G} = ((M_i)_{i\in I}, g)$, where M_i is the set of actions for agent i and $g\colon M \times \Theta \to \mathbb{R}^I$ assigns a payoff profile to each pair of action profile and the external state. We say that $\mathcal{G}$ is a *V-game* if $g(m,\theta) \in V$ for every $m \in M$ and every $\theta \in \Theta$. Every contingent mechanism $\mathcal{M} = ((M_i)_{i\in I}, O)$ with $O\colon M \times \Theta \to \Delta(Z)$ induces a V-game $\mathcal{G} = ((M_i)_{i\in I}, g)$ with

$$g(m,\theta) = \left(\sum_{z\in Z} u_i(z)\, O(m,\theta)(z)\right)_{i\in I}$$

for each $m \in M$ and each $\theta \in \Theta$; conversely, every V-game can be induced by some contingent mechanism.

Our definition of ICR in this case corresponds exactly to that in DFM. Our Theorem 1 now proves that two types have the same belief hierarchy over Θ if and only if they have the same ICR actions in all V-games. In the case that V is a non-degenerate product set, i.e.,

$$V = \prod_{i\in I} \left[\underline{v}_i, \overline{v}_i\right]$$

with $\underline{v}_i < \overline{v}_i$ for every $i \in I$, this result was already proved in DFM. Specifically, for every non-degenerate product set V, Dekel et al. (2006, Lemma 4) show that if two types have distinct belief hierarchies, then there is a V-game where they have disjoint ICR action sets;[17] conversely, Dekel et al. (2007, Proposition 1 and Corollary 2) show that two types with the same belief hierarchy have the same ICR actions (for finite types and general types, respectively) in every V-game.

[16] There is an alternative interpretation of DFM under which their results can be seen as a special case of the results in this paper without appeal to "external" states. Observe that uncontingent mechanisms and private states – profiles of extremal preferences, in our simplex representation – jointly define a set of utility functions from message profiles and states to payoffs, i.e., a game with incomplete information over $(U_i)_{i\in I}$. If the outcome space were sufficiently rich, this problem would reduce to a version of DFM. If not, results in this paper would identify strategic distinguishability in restricted classes of games.

[17] Dekel et al. (2006, Lemma 4) prove something a little stronger: for every distance between nth order beliefs, we can find $\varepsilon > 0$ such that no action is both δ-interim correlated rationalizable for one type and $(\delta + \varepsilon)$-interim correlated rationalizable for the other type.

The assumption that the set V is a non-degenerate product set has a natural counterpart in our setting. Say that we have a *private good environment* if the outcome space Z has a product structure $Z = \prod_{i \in I} Z_i$, and each agent i's utility from outcome z depends only on the ith component z_i, so $u_i(z) = \widetilde{u}_i(z_i)$ for some $\widetilde{u}_i : Z_i \to \mathbb{R}$. In this case, the set of feasible payoff vectors has the product structure

$$V = \prod_{i \in I} \left[\underline{v}_i, \overline{v}_i\right],$$

where

$$\underline{v}_i = \min_{z_i \in Z_i} \widetilde{u}_i(z_i) \text{ and } \overline{v}_i = \max_{z_i \in Z_i} \widetilde{u}_i(z_i).$$

But our Theorem 1 did not rely on the private good environment assumption. If the assumption of common certainty of conditional preferences is maintained but the private good assumption is dropped, then the set V of feasible payoff profiles could be any convex polytope whose projection in each dimension is non-degenerate. For example, our Theorem would apply to environments where

$$V = \left\{ v \in [-1, 1]^I \,\middle|\, \sum_{i \in I} v_i = 0 \right\}$$

so we restricted attention to zero sum games. And it would apply to environments where

$$V = \left\{ v \in [0, 1]^I \,\middle|\, v_i = v_j \text{ for all } i, j \in I \right\},$$

so we restricted attention to common interest games. Thus, while the original proof of Dekel et al. (2006, Lemma 4) relied on the assumption that all payoff vectors are feasible, our Theorem 1 – with external states added and common certainty of conditional preferences assumed – establishes that it would remain true if DFM had restricted attention to zero sum games, common interest games, or many other subsets of games which restricted how agents' payoffs can vary.

Gossner and Mertens (2001) show that a zero sum Bayesian game has a value which depends only on the probability distribution over Mertens–Zamir hierarchies and is increasing in informativeness in Blackwell's sense. The argument requires a strategic distinguishability result for the case of zero sum games.[18] While the formulation of our strategic distinguishability question and the proof are different from those arising in Gossner and Mertens (2001), the argument above suggests when and how the approach in this paper could be used to develop analogous strategic distinguishability exercises in different classes of games.

5.3. *Strategic distinguishability without simplex representations*

The simplex representation was convenient in stating and proving our results, and relating them to the existing literature. We will now state our main result without reference to a simplex representation. A cost of doing so is that we lose our utility representations of interdependent preferences. We do so nonetheless in order to verify the independence of the result from the simplex representation chosen, and also as a prelude to relaxing our uniform ranking and bounded utility assumptions in the next section, where a simplex representation is not available.

[18] Gossner and Mertens (2001) is an abstract of unpublished work; we are grateful to Olivier Gossner for privately sharing notes from the complete paper.

Let X be a countable set of states.[19] Recall that $F(X)$ denotes the set of all acts over X. Let $P(X)$ be the set of all preferences $\succsim$ over $F(X)$ represented by a belief about states $\mu \in \Delta(X)$ and a μ-absolutely summable state-dependent utility index $u\colon X \times Z \to \mathbb{R}$, i.e., $\sum_x |u(x,z)|\mu(x) < \infty$ for every $z \in Z$, as follows:

$$f \succsim f' \Leftrightarrow \sum_{x\in X}\sum_{z\in Z} u(x,z)(f(x)(z) - f'(x)(z))\mu(x) \geq 0.$$

With this notation, we can define a countable type space $\mathcal{T} = (T_i, \pi_i)_{i\in I}$ with

$$\pi_i\colon T_i \to P(T_{-i}),$$

where for each type $t_i \in T_i$, $\pi_i(t_i)$ denotes the preference of type t_i over acts over the opponents' types. We write $\hat{\pi}_{i,1}(t_i)$ for the restriction of $\pi_i(t_i)$ to lotteries, and call it his first order preference. We also write $\hat{\pi}_{i,2}(t_i)$ for the restriction of $\pi_i(t_i)$ to acts that depend only on the opponents' first order preferences, and we call it his second order preference. We define third order, and higher order, preferences similarly, and we write $\hat{\pi}_i(t_i) = (\hat{\pi}_{i,1}(t_i), \hat{\pi}_{i,2}(t_i), \ldots)$ for the hierarchy of type t_i's higher order preferences.

We now impose the uniform ranking and bounded utility assumptions on preferences. Given a pair of outcomes $\overline{z}, \underline{z} \in Z$ and a utility bound $B \geq 1$, we say that a preference $\succsim \in P(X)$ is $(\overline{z}, \underline{z}, B)$-*bounded* if it is represented by (μ, u) such that

1. uniform ranking: $u(x,\overline{z}) > u(x,\underline{z})$ for every $x \in X$; we normalize each $u(x,\cdot)$ so that $u(x,\overline{z}) = 1$ and $u(x,\underline{z}) = 0$;
2. bounded utility: $|u(x,z)| \leq B$ for every $z \in Z$ (given the above normalization).

Let $P_{\overline{z},\underline{z},B}(X)$ be the set of all $(\overline{z}, \underline{z}, B)$-bounded preferences over $F(X)$. Given a profile $\mathcal{B} = (\overline{z}_i, \underline{z}_i, B_i)_{i\in I}$ of pairs of outcomes $\overline{z}_i, \underline{z}_i \in Z$ and utility bounds $B_i \geq 1$, we say that a countable type space $\mathcal{T} = (T_i, \pi_i)_{i\in I}$ is $\mathcal{B}$-*bounded* if $\pi_i(t_i) \in P_{\overline{z}_i,\underline{z}_i,B_i}(T_{-i})$ for every $i \in I$ and $t_i \in T_i$.

Given a countable type space $\mathcal{T} = (T_i, \pi_i)_{i\in I}$ and a finite mechanism $\mathcal{M} = ((M_i)_{i\in I}, O)$, we say that a profile $\sigma = (\sigma_i)_{i\in I}$ of behavioral strategies $\sigma_i\colon T_i \to \Delta(M_i)$ is an *equilibrium* if $\pi_i(t_i)$ weakly prefers $O(m_i,\cdot)\circ\sigma_{-i}$ to $O(m_i',\cdot)\circ\sigma_{-i}$ for every agent $i \in I$, every type $t_i \in T_i$, and every messages $m_i, m_i' \in M_i$ with $\sigma_i(t_i)(m_i) > 0$.[20] Let $E_i(t_i)$ denote the set of actions played by type t_i with positive probability in some equilibrium. Given $\mathcal{B} = (\overline{z}_i, \underline{z}_i, B_i)_{i\in I}$, we also define the set of actions that are $\mathcal{B}$-*boundedly rationalizable for type* t_i, denoted by $R_{i,\mathcal{B}}(t_i)$, as follows:

$$R^0_{i,\mathcal{B}}(t_i) = M_i,$$

$$R^{n+1}_{i,\mathcal{B}}(t_i) = \left\{ m_i \in M_i \;\middle|\; \begin{array}{l} \text{there exists } \succsim_i \in P_{\overline{z}_i,\underline{z}_i,B_i}(M_{-i}\times T_{-i}) \text{ s.t.} \\ \text{(i) } \succsim_i \text{ is certain of } \prod_{j\neq i} \text{graph}(R^n_{j,\mathcal{B}}), \\ \text{(ii) } \text{mrg}_{T_{-i}} \succsim_i = \pi_i(t_i), \\ \text{(iii) } \succsim_i \text{ weakly prefers } O(m_i,\cdot) \text{ to } O(m_i',\cdot) \text{ for every } m_i' \in M_i \end{array} \right\},$$

$$R_{i,\mathcal{B}}(t_i) = \bigcap_{n=0}^{\infty} R^n_{i,\mathcal{B}}(t_i),$$

[19] We assume countable state spaces to avoid measurability issues as well as to guarantee the existence of equilibria.

[20] We define $O(m_i,\cdot)\circ\sigma_{-i}$ as an act over T_{-i} given by $O(m_i,\sigma_{-i})(t_{-i})(z) = \sum_{m_{-i}} O(m_i,m_{-i})(z) \times \prod_{j\neq i}\sigma_j(t_j)(m_j)$ for every $t_{-i} \in T_{-i}$ and $z \in Z$.

where we say that $\succsim \in P(X)$ is *certain of* $E \subseteq X$ if $X \setminus E$ is Savage-null with respect to $\succsim$, i.e., $f \sim f'$ whenever f and f' agree on E, and the *marginal of* $\succsim \in P(X \times Y)$ *on* X, denoted by $\text{mrg}_X \succsim \in P(X)$, is the restriction of $\succsim$ to $F(X)$.

Then we have the following result.

Theorem 3. *Fix a profile of uniform rankings and utility bounds $\mathcal{B} = (\overline{z}_i, \underline{z}_i, B_i)_{i \in I}$. For every pair of countable and $\mathcal{B}$-bounded type spaces $\mathcal{T} = (T_i, \pi_i)_{i \in I}$ and $\mathcal{T}' = (T'_i, \pi'_i)_{i \in I}$, every agent $i \in I$, and every pair of types $t_i \in T_i$ and $t'_i \in T'_i$, the following three conditions are equivalent:*

1. $\hat{\pi}_i(t_i; \mathcal{T}) = \hat{\pi}_i(t'_i; \mathcal{T}')$;
2. $R_{i,\mathcal{B}}(t_i; \mathcal{T}, \mathcal{M}) \cap R_{i,\mathcal{B}}(t'_i; \mathcal{T}', \mathcal{M}) \neq \emptyset$ *for every finite mechanism* $\mathcal{M}$;
3. $E_i(t_i; \mathcal{T}, \mathcal{M}) \cap E_i(t'_i; \mathcal{T}', \mathcal{M}) \neq \emptyset$ *for every finite mechanism* $\mathcal{M}$.

Proof. For each $i \in I$, let U_i be the set of extreme points of a simplex such that $\text{co}(U_i)$ contains all utility indices u_i with $u_i(\overline{z}_i) = 1$, $u_i(\underline{z}_i)$, and $|u_i(z)| \leq B_i$ for every $z \in Z$. Note that two types have the same hierarchy of interdependent preferences if and only if they can be represented in $(U_i)_{i \in I}$-based type spaces with the same hierarchy of beliefs over $(U_i)_{i \in I}$. $1 \Rightarrow 3$ follows from Theorem 2 and this fact. $3 \Rightarrow 2$ follows from the fact that equilibrium is a refinement of $\mathcal{B}$-bounded rationalizability. $2 \Rightarrow 1$ follows from Theorem 1, the previous fact, and the fact that $\mathcal{B}$-bounded rationalizability is a refinement of ICR based on $(U_i)_{i \in I}$. □

Theorem 3 is a simple rewriting of Theorems 1 and 2.

This statement of the theorem addresses the following two issues that arise from our modeling choice.

Too rich type spaces First, in Section 2, among other universal type spaces, we discussed a payoff universal type space T_P for our conditional altruism example, but dismissed it as it was too rich for two distinct types to be strategically distinguishable. Indeed, if we adopted too rich a universal type space, two different universal types may not be strategically distinguishable. To be specific, let us arbitrarily pick a profile of finite sets of utility indices $\Omega_i \subset \text{co}(U_i)$ and an $(\Omega_i)_{i \in I}$-based type space $\mathcal{T}_\Omega = (T_i, \mu_i)_{i \in I}$ with $\mu_i : T_i \to \Delta(\Omega_i \times T_i)$. If Ω_i is not the set of extreme points of a simplex, then a point in $\text{co}(\Omega_i)$ may be represented by two different convex combinations of Ω_i, and hence two types with different hierarchies of beliefs over $(\Omega_i)_{i \in I}$ may have the same hierarchy of interdependent preferences. By Theorem 3, two countable types are strategically distinguishable if and only if they have different hierarchies of interdependent preferences. Thus, we cannot strategically distinguish two types with the same hierarchy of interdependent preferences even if they have different hierarchies of beliefs over $(\Omega_i)_{i \in I}$.

Change of "coordinate systems" Second, we may be able to represent the same type space in two different ways, $(U_i)_{i \in I}$-based and $(U'_i)_{i \in I}$-based type spaces, where both are the sets of extreme points of simplices. Clearly, the same hierarchy of interdependent preferences can be represented differently by hierarchies of beliefs over $(U_i)_{i \in I}$ and over $(U'_i)_{i \in I}$, which, in turn, can have different versions of ICR. However, although ICR depends on the simplex profile, equilibrium does not. Thus strategic distinguishability for equilibrium, whether a pair of types

have disjoint equilibrium action sets or not, does not depend on the simplex profile.[21] Moreover, two types have the same hierarchy of beliefs over $(U_i)_{i\in I}$ if and only if they have the same hierarchy of beliefs over $(U_i')_{i\in I}$. In this sense, the choice of the simplex profile is irrelevant for our characterization that identifies which pair of types have the same hierarchy of beliefs over $(U_i)_{i\in I}$.

5.4. Relaxing the uniform ranking and bounded utility assumptions

Here, we show that it is possible to relax the uniform ranking assumption and assume instead that preferences are not completely indifferent over outcomes.[22] At the same time, we also replace the bounded utility assumption by what we call the "λ-continuity" assumption, which imposes conditions on preferences directly.

Given $\lambda \in (0, 1/2]$, we say that a preference $\succsim \in P(X)$ is λ-*continuous* if

1. no complete indifference over outcomes: there exists a pair of outcomes $\overline{z}, \underline{z} \in Z$ such that $\overline{z} \succ \underline{z}$;
2. λ-continuity: $(1-\lambda)\overline{z} + \lambda f \succsim (1-\lambda)\underline{z} + \lambda f'$ for every $f, f' \in F(X)$ (given the above pair $\overline{z}, \underline{z}$).

Thus we require that $\succsim$ not be completely indifferent over outcomes, and that the preference relation $\overline{z} \succ \underline{z}$ be maintained at least weakly even if we mix these outcomes $\overline{z}$ and $\underline{z}$ with a small probability of arbitrary acts f and f', respectively. Thus λ-continuity imposes a bound on the state sensitivity of preferences. In terms of expected utility representations, a preference $\succsim \in P(X)$ represented by $\mu \in \Delta(X)$ and $u\colon X \times Z \to \mathbb{R}$ is λ-continuous if and only if

$$\max_{z,z'\in Z} \sum_{x\in X} (u(x,z) - u(x,z'))\mu(x) > 0,$$

$$\sum_{x\in X} \max_{z,z'\in Z} (u(x,z) - u(x,z'))\mu(x) \leq \frac{1-\lambda}{\lambda} \max_{z,z'\in Z} \sum_{x\in X} (u(x,z) - u(x,z'))\mu(x).$$

Therefore, λ-continuity for sufficiently small λ is a weakening of the bounded utility assumption. To see this, if a preference is $(\overline{z}, \underline{z}, B)$-bounded, then with our normalization of utility indices, the left-hand side of the second inequality is bounded above by $2B$, while the right-hand side is bounded below by $(1-\lambda)/\lambda$. Thus the inequality holds as long as $\lambda \leq 1/(2B+1)$.

Let $P_\lambda(X)$ be the set of all λ-continuous preferences over $F(X)$. A countable type space $\mathcal{T} = (T_i, \pi_i)_{i\in I}$ is λ-*continuous* if $\pi_i(t_i) \in P_\lambda(T_{-i})$ for every $i \in I$ and $t_i \in T_i$. We define the set of actions that are *B-boundedly rationalizable for type* t_i, denoted by $R_{i,\lambda}(t_i)$, similarly to $R_{i,\mathcal{B}}(t_i)$; we simply replace $P_{\overline{z}_i,\underline{z}_i,B_i}$ by P_λ.

We can now state our strategic distinguishability results for λ-continuous preferences.

[21] In fact, strategic distinguishability for ICR does not depend on the simplex profile, either. Indeed, take two types t_i and t_i' from $(U_i)_{i\in I}$-based type spaces $(T_i, \mu_i)_{i\in I}$ and $(T_i', \mu_i')_{i\in I}$ with the same hierarchy of beliefs over $(U_i)_{i\in I}$. Then they have the same ICR with private state spaces $(U_i)_{i\in I}$. By changing "coordinate systems", we consider $(U_i')_{i\in I}$-based type spaces $(T_i, \mu_i'')_{i\in I}$ and $(T_i', \mu_i''')_{i\in I}$. Then t_i and t_i' have the same hierarchy of beliefs over $(U_i')_{i\in I}$, and hence have the same ICR with private state spaces $(U_i')_{i\in I}$ as well although this ICR may be different from ICR with $(U_i)_{i\in I}$.

[22] The absence of complete indifference is a maintained assumption in the virtual and Bayesian implementation literature, e.g., Abreu and Sen (1991) and Duggan (1997) as well as Abreu and Matsushima (1992).

Theorem 4. *Fix $\lambda \in (0, 1/2]$. For every pair of countable and λ-continuous type spaces $\mathcal{T} = (T_i, \pi_i)_{i \in I}$ and $\mathcal{T}' = (T_i', \pi_i')_{i \in I}$, every agent $i \in I$, and every pair of types $t_i \in T_i$ and $t_i' \in T_i'$, the following three conditions are equivalent:*

1. $\hat{\pi}_i(t_i; \mathcal{T}) = \hat{\pi}_i(t_i'; \mathcal{T}')$;
2. $R_{i,\lambda}(t_i; \mathcal{T}, \mathcal{M}) \cap R_{i,\lambda}(t_i'; \mathcal{T}', \mathcal{M}) \neq \emptyset$ *for every finite mechanism* $\mathcal{M}$;
3. $E_i(t_i; \mathcal{T}, \mathcal{M}) \cap E_i(t_i'; \mathcal{T}', \mathcal{M}) \neq \emptyset$ *for every finite mechanism* $\mathcal{M}$.

The proof is in Appendix C. Note that our strategic distinguishability results hold for equilibrium, λ-continuous rationalizability and everything in between. Thus, Lemma 2 of Abreu and Matsushima (1992) is a special case of Theorem 4 since every finite type space with no complete indifference over outcomes is λ-continuous with some $\lambda > 0$, and for this value of λ, their solution concept of iterated elimination of strictly dominated actions is in between equilibrium and λ-continuous rationalizability.[23]

5.5. Relaxing the λ-continuity assumption

A type of an agent with completely indifferent preferences cannot be distinguish from an agent with any other preferences, so the no complete indifference assumption must be maintained. But what can we say if we drop the bounded utility or λ-continuity assumption altogether? The following result will continue to be true.

Proposition 4. *For every pair of countable type spaces $\mathcal{T} = (T_i, \pi_i)_{i \in I}$ and $\mathcal{T}' = (T_i', \pi_i')_{i \in I}$, every agent $i \in I$, and every pair of types $t_i \in T_i$ and $t_i' \in T_i'$, we have*

$$\hat{\pi}_i(t_i; \mathcal{T}) = \hat{\pi}_i(t_i'; \mathcal{T}') \Rightarrow E_i(t_i; \mathcal{T}, \mathcal{M}) \cap E_i(t_i'; \mathcal{T}', \mathcal{M}) \neq \emptyset$$

for every finite mechanism $\mathcal{M}$.

However, Proposition 2 does not extend without any a priori bound on utilities. Note that we are not asking whether two types can be strategically distinguished or not; we are asking whether strategic distinguishability is characterized merely by interdependent preference hierarchies. And the latter requires distinguishing two *sets* of types, each of which corresponds to an interdependent preference hierarchy. In Proposition 2, we constructed a finite mechanism, depending only on $\varepsilon > 0$ and simplex profile $(U_i)_{i \in I}$, that can strategically distinguish two sets of types as long as the two sets correspond to two belief hierarchies over $(U_i)_{i \in I}$ that are ε apart from each other. Here, we will show that without the bounded utility or λ-continuity assumption, there is no finite mechanism that can strategically distinguish two sets of types that correspond to two interdepen-

[23] There are two technical differences between AM's formulation and ours. As noted in footnote 2, AM allow for all simple (i.e., finite support) lotteries over any (possibly infinite) set of outcomes. AM show that if we focus on finite (or "regular") mechanisms, and rule out mechanisms that involve integer games, then a social choice function that is virtually implementable in mixed-strategy equilibrium must satisfy the measurability condition. Duggan (1997) provides an example of a social choice function that is not measurable, but can be exactly implemented in pure-strategy equilibrium by a finite mechanism. Serrano and Vohra (2010) extend Duggan's argument and show that the social choice function is indeed exactly implementable in mixed-strategy equilibrium by an infinite mechanism.

dent preference hierarchies with the same first order preference (even if they differ in the second and higher order preferences).[24]

Proposition 5. *For every pair of interdependent preference hierarchies of finite types* $h = (\succsim_1, \succsim_2, \ldots)$, $h' = (\succsim'_1, \succsim'_2, \ldots)$ *such that* $\succsim_1 = \succsim'_1$, *every agent* $i \in I$, *and every finite mechanism* $\mathcal{M} = ((M_i)_{i\in I}, O)$, *there exist a pair of finite type spaces* $\mathcal{T} = (T_i, \mu_i, u_i)_{i\in I}$ *and* $\mathcal{T}' = (T'_i, \mu'_i, u'_i)_{i\in I}$ *and a pair of types* $t_i \in T_i$ *and* $t'_i \in T'_i$ *such that* $\hat{\pi}_i(t_i; \mathcal{T}) = h$, $\hat{\pi}_i(t'_i; \mathcal{T}') = h'$, *and* $E_i(t_i; \mathcal{T}, \mathcal{M}) = E_i(t'_i; \mathcal{T}', \mathcal{M})$.

The proof is in Appendix C. In what follows, we use an example to illustrate the difficulty of strategically distinguishing higher order preferences without restrictions.

In what follows, we use an example to illustrate why a mechanism like the one constructed in Section 4.2 cannot strategically distinguish types with distinct hierarchies of interdependent preferences. Take the conditional altruism example, and consider a mechanism with two messages 0 and 1 for each agent and the outcome being in the form of

$$O(m_1, m_2)(z) = (1-\varepsilon)O_1(m_1)(z) + \varepsilon O_2(m_1, m_2)(z)$$

with $\varepsilon \geq 0$, where O_1 is to solicit agent 1's report about his first order preference, whereas O_2 is to solicit both agents' reports about their higher order preferences. To fix ideas, suppose that $O_1(m_1 = 0)$ gives the prize to nobody and agent 1 with probability $1/2$ each, and $O_1(m_1 = 1)$ gives the prize to agent 2; $O_2(m_1, m_2)$ gives the prize to agents 1 and 2 with probability $m_1 m_2/2$, and to nobody with the remaining probability $1 - m_1 m_2$. Consider a type space, where each agent i has two possible types 0 and 1, each type believes that the opponent's type is 0 or 1 with probability $1/2$, and payoff parameters (the payoff from the opponent getting the prize) are given by

	$t_2 = 0$	1
$t_1 = 0$	$1+v, 1+v$	$1-v, 1$
1	$1, 1-v$	$1, 1$

with $v \in \mathbb{R}$. Note that all types have the same expected value of the payoff parameter $(1+v)/2 + (1-v)/2 = 1$, and hence have the same interdependent preference hierarchy as the truly altruistic type with complete information, independently of v.

In this case, if $\varepsilon = 0$, then agent 1 has an incentive to report $m_1 = 1$ (as a dominant action) according to his first order preference. But since there is no interaction term between m_1 and m_2, no information about higher order preferences can be revealed in equilibrium actions. In contrast, if $\varepsilon > 0$, then for sufficiently large v, type 0 of agent 1 no longer has a dominant action, and indeed, the strategy profile of reporting $m_i = t_i$ becomes an equilibrium. In sum, there is no $\varepsilon \geq 0$ that keeps agent 1's incentive to report his first order preference truthfully and yet solicits higher order preferences from both agents.

Note that this example hinges crucially on the difference between us and AM: our exercise of strategic distinguishability (Theorems 1 and 2) is to construct a mechanism independently of an underlying type space, whereas AM fix a finite type space first and then construct a mechanism. This example also illustrates a trade-off between ε and v. There is no $\varepsilon > 0$ that keeps $m_1 = 1$ a dominant action for agent 1 independently of v. But if we knew a bound on v, then we could

[24] Proposition 5 is stated and proved for finite types, which suffices to show the impossibility of strategic distinguishability, but would continue to hold with countable types.

choose ε small enough (in the magnitude of $1/|v|$) so that $m_1 = 1$ is a dominant action for agent 1.

6. Discussion

6.1. Strategic equivalence

For any interim solution concept, we say that two types are *strategically equivalent* if they have the same set of solutions for every finite mechanism. It follows from Proposition 1 and Theorem 1 that hierarchies of beliefs over $(U_i)_{i\in I}$ characterize strategic equivalence for ICR. However, the corresponding result does not hold for equilibrium even if we restrict attention to countable type spaces. It is well known from the setting with common certainty of conditional preferences that solution concepts such as equilibrium and interim independent rationalizability depend on redundant types. See Dekel et al. (2007), Ely and Peski (2006) and Sadzik (2010).

6.2. Separating beliefs from payoffs

Our approach integrates the treatment of payoffs (or utility indices that represent conditional preferences over lotteries) with beliefs and higher order beliefs about those payoffs. But the standard approach in the literature has been to discuss the two separately. In particular, as we discussed in Section 2, an alternative construction is to first identify a general space of interdependent payoff types, such as that of Gul and Pesendorfer (2016), and then allow for all possible beliefs and higher order beliefs over those payoff types. But this construction gives a space of interdependent types different from our space of strategically distinguishable types. In particular, as we discussed in Section 2, this alternative construction is not tight because it will label types differently even if they differ only in what their conditional preferences would be given zero probability events. On the other hand, it is not rich enough because it does not allow conditional preferences to depend on others' beliefs. For example, I might be more altruistic if I believe that you believe that I am altruistic. This cannot arise in the Gul and Pesendorfer (2016) construction, where payoff types depend only on others' payoff types, not their beliefs. Thus, we allow conditional preferences to depend on beliefs, as in the "psychological games" literature of Geanakoplos et al. (1989) and Battigalli and Dufwenberg (2009). However, preferences depend only on other agents' beliefs about others' types and not – as in the psychological games literature – on beliefs about actions.[25]

However, even though our space is quite different from this alternative construction, it still makes sense to ask if and how we can distinguish between "payoff types" and "belief types" in a natural way in our space. Just as beliefs cannot be pinned down in (single person) expected utility representations of preferences unless we fix a numéraire, there is indeterminacy in beliefs in our construction based on the choice of representations of the extreme preferences U_i. But particular applications may suggest a numéraire over which the modeler wishes to treat utility as state independent, which will pin down the representation. A type of agent i can then be characterized by a belief over others' types, and conditional preferences over lotteries given others' types. We can use the separation to interpret existing works.

[25] Such beliefs can be captured as "characteristics" in Gul and Pesendorfer (2016).

6.3. Operational meaning and revealed preference

Two types are strategically distinguishable if and only if there exists a finite mechanism where they are guaranteed to behave differently. No additional information is required to identify the interdependent hierarchies. As we discussed in Sections 5.1 and 5.2, if mechanisms could depend on some external states, it would be possible to learn about agents' beliefs are higher order beliefs about external states. If mechanisms cannot depend on any additional data, the interdependent hierarchies are all that can be operationally identified, and we cannot distinguish between psychological or informational origins of the interdependence.

Classical single person revealed preference theory characterizes when a set of choice functions are consistent with rational choice (Afriat, 1967), with the weak axiom of revealed preference (WARP) being the key restriction on choice rules. If, in addition to standard rationality assumptions, we looked at choices over lotteries and added the independence assumption, we would obtain more restrictions. A primitive single person revealed preference question would then be if you can tell the difference between two different expected utility preferences over lotteries. A standard argument says that we can construct a pair of lotteries such that one preference will lead to one strict ordering, and the other preference will lead to the opposite strict ordering. Our strategic distinguishability question is a many person analogue of this revealed preference question.[26]

6.4. The expected utility assumption

We maintained the assumption of expected utility maximization, but dispensed with monotonicity to incorporate the interdependence of preferences we want to capture. Epstein and Wang (1996) construct a universal type space of non-expected utility preferences, incorporating non-expected utility preferences such as ambiguity aversion, but maintaining monotonicity as well as additional regularity conditions. Di Tillio (2008) allows general preferences, and thus does not require monotonicity or independence, but restricts attention to preferences over finite outcomes at every order of the hierarchy.[27]

Acknowledgments

The first two authors acknowledge financial support through NSF Grants SES 0851200 and 1215808. We are grateful for valuable and detailed comments from the editor, Marciano Siniscalchi, an associate editor and two referees, as well as seminar/conference participants at Bocconi, Columbia, Chicago, Harvard/MIT, HEC, Kyoto, Northwestern, NYU, Oxford, Penn, SAET, Yale, Warwick and the Econometric Society World Congress in Shanghai. We acknowledge valuable research assistance from Áron Tóbiás.

[26] There is a small literature developing strategic analogues of classic single agent decision theory. See, for example, Sprumont (2000). There are many differences between this paper and that literature. Thus, Sprumont (2000) fixes the action set while we allow for arbitrary action sets. His theory is ordinal and does not impose expected utility while ours is cardinal and does impose expected utility.

[27] The strategic distinguishability question does not appear to have been addressed without expected utility preferences. Chambers (2008) shows the impossibility of constructing a uniform scoring rule to distinguish preferences and beliefs in a non-expected utility setting, which suggests that positive results about strategic distinguishability would be hard to obtain. Grant et al. (2016) analyze "Savage games" played on subjective state spaces, allowing both expected utility maximizers and more general preferences; in neither case do they consider strategic distinguishability.

Appendix A. Proofs in Section 4

A.1. Proof of Lemma 1

Suppose not. Then there exists $\varepsilon > 0$ such that for every $n \in \mathbb{N}$, there exist $\mu_n, m_n, m'_n \in \Delta(U \times X)$ with $d_\Delta(\mu_n, m_n) \leq 1/n$ and $d_\Delta(\mu_n, m'_n) > \varepsilon$, measurable space Ω_n, perturbed outcome function $O_n \colon M^0 \times X \times \Omega_n \to \Delta(Z)$ with $\|O_n(\cdot,\cdot,\omega) - O^0\| \leq 1/n$ for every $\omega \in \Omega_n$, $\mu'_n \in \Delta(U \times X \times X' \times \Omega_n)$ with $X' = X$ such that $\mu'_n\left(U \times \left\{(x,x') \mid d(x,x') \leq \delta\right\} \times \Omega_n\right) = 1$, $\mathrm{mrg}_{U\times X}\mu'_n = \mu_n$, and $\mathrm{mrg}_{U\times X'\times\Omega}\mu'_n$ weakly prefers $O_n(m'_n,\cdot,\cdot)$ to $O_n(m_n,\cdot,\cdot)$. Since X is a compact metric space, by taking a subsequence if necessary, we can find $\mu^*, m'^{,*} \in \Delta(U \times X)$ such that $\mu_n \to \mu^*$ and $m'_n \to m'^{,*}$ as $n \to \infty$. Note that $m_n \to \mu^*$ as $n \to \infty$, and $\mu^* \neq m'^{,*}$. Let

$$u^* = \int \sum_z u(z) O^0(\mu^*, x)(z) d\mu^*(u,x).$$

Claim 1. *We have*

$$\lim_{n\to\infty} \int \sum_z u(z) O^0(m_n, x)(z) d\mu_n(u,x) = u^*,$$

$$\limsup_{n\to\infty} \int \sum_z u(z) O^0(m'_n, x)(z) d\mu_n(u,x) < u^*.$$

Proof of Claim 1. The claim follows from showing that

$$\lim_{n\to\infty} \int \sum_z u(z) C_{m_n}(f_k, f_l)(x)(z) d\mu_n(u,x) = \int \sum_z u(z) C_{\mu^*}(f_k, f_l)(x)(z) d\mu^*(u,x),$$

$$\limsup_{n\to\infty} \int \sum_z u(z) C_{m'_n}(f_k, f_l)(x)(z) d\mu_n(u,x) \leq \int \sum_z u(z) C_{\mu^*}(f_k, f_l)(x)(z) d\mu^*(u,x),$$

for each k, l, and that the second inequality holds with strict inequality for some k, l. The first equality and the second weak inequality follow from the standard revealed preference argument. To show the strict inequality, since $\mu^* \neq m'^{,*}$ and $F \subset F_c(X)$ is dense in the sup norm, there exist k, l such that μ^* strictly prefers f_k to f_l while $m'^{,*}$ strictly prefers f_l to f_k. Since m'_n strictly prefers f_l to f_k for sufficiently large n, we have:

$$\begin{aligned}
&\lim_{n\to\infty} \int \sum_z u(z) C_{m'_n}(f_k, f_l)(x)(z) d\mu_n(u,x) \\
&= \lim_{n\to\infty} \int \sum_z u(z) f_l(x)(z) d\mu_n(u,x) \\
&= \int \sum_z u(z) f_l(x)(z) d\mu^*(u,x) \\
&< \int \sum_z u(z) f_k(x)(z) d\mu^*(u,x) \\
&= \int \sum_z u(z) C_{\mu^*}(f_k, f_l)(x)(z) d\mu^*(u,x).
\end{aligned}$$

which establishes the claim. □

Claim 2. *We have*

$$\lim_{n\to\infty}\left(\int\sum_z u(z)O_n(m,x',\omega)(z)d\mu'_n(u,x,x',\omega) - \int\sum_z u(z)O^0(m,x)(z)d\mu_n(u,x)\right)=0$$

and the convergence is uniform in $m\in M^0$.

Proof of Claim 2. Note that

$$\begin{aligned}
&\left|\int\sum_z u(z)O_n(m,x',\omega)(z)d\mu'_n(u,x,x',\omega) - \int\sum_z u(z)O^0(m,x)(z)d\mu_n(u,x)\right|\\
&\le\left|\int\sum_z u(z)O_n(m,x',\omega)(z)d\mu'_n(u,x,x',\omega)\right.\\
&\quad\left.-\int\sum_z u(z)O^0(m,x')(z)d\mu'_n(u,x,x',\omega)\right|\\
&\quad+\left|\int\sum_z u(z)O^0(m,x')(z)d\mu'_n(u,x,x',\omega) - \int\sum_z u(z)O^0(m,x)(z)d\mu_n(u,x)\right|.
\end{aligned}$$

The first term is bounded above by $(1/n)\max_{u,z,z'}|u(z)-u(z')|$ since $\|O_n(\cdot,\cdot,\omega)-O^0\|\le 1/n$ for every $\omega\in\Omega_n$.

To show that the second term converges to 0 uniformly in m, it is enough to show that

$$\lim_{n\to\infty}\left(\int\sum_z u(z)f(x')(z)d\mu'_n(u,x,x',\omega) - \int\sum_z u(z)f(x)(z)d\mu_n(u,x)\right)=0$$

for each $f\in F_c(X)$. Since X is a compact metric space, f is uniformly continuous. Therefore, for every $\eta>0$, there exists N such that $\max_z\left|f(x)(z)-f(x')(z)\right|<\eta$ whenever $d(x,x')\le 1/N$. For every $n\ge N$, we have

$$\begin{aligned}
&\left|\int\sum_z u(z)f(x')(z)d\mu'_n(u,x,x',\omega) - \int\sum_z u(z)f(x)(z)d\mu_n(u,x)\right|\\
&\le\left|\int\sum_z u(z)f(x')(z)d\mu'_n(u,x,x',\omega) - \int\sum_z u(z)f(x)(z)d\mu'_n(u,x,x',\omega)\right|\\
&\quad+\left|\int\sum_z u(z)f(x)(z)d\mu'_n(u,x,x',\omega) - \int\sum_z u(z)f(x)(z)d\mu_n(u,x)\right|.
\end{aligned}$$

The first term is bounded above by $\eta\max_{u,z,z'}|u(z)-u(z')|$; the second term is equal to zero since $\mathrm{mrg}_{U\times X}\mu'_n=\mu_n$. □

We can now complete the proof of Lemma 1 since Claims 1 and 2 contradict the assumption that $\mathrm{mrg}_{U\times X'\times\Omega}\mu'_n$ weakly prefers $O_n(m'_n,\cdot,\cdot)$ to $O_n(m_n,\cdot,\cdot)$.

A.2. Proof of Lemma 2

The proof is by induction on n. Suppose that for each $k \leq n-1$, $m_i \in R_i^{n-1}(t_i; \mathcal{T}, \mathcal{M})$ implies $d_{i,k}(\hat{\mu}_{i,k}(t_i), m_{i,k}) \leq \varepsilon_k \leq \varepsilon_{n-1}$ for every agent $i \in I$ and every type $t_i \in T_i$. Suppose that there exists $m_i^* \in R_i^n(t_i; \mathcal{T}, \mathcal{M})$ such that $d_{i,n}(\hat{\mu}_{i,n}(t_i), m_{i,n}^*) > \varepsilon_n$. Then there exists $\nu_i \in \Delta(M_{-i} \times U_i \times T_{-i})$ such that $\nu_i(\{(m_{-i}, u_i, t_{-i}) \mid m_{-i} \in R_{-i}^{n-1}(t_{-i}; \mathcal{T}, \mathcal{M})\}) = 1$, $\mathrm{mrg}_{U_i \times T_{-i}} \nu_i = \mu_i(t_i)$, and ν_i weakly prefers $O(m_i^*, \cdot)$ to $O(m_i', \cdot)$ for every $m_i' \in M_i$.

Collect all the terms in O that depend on $m_{i,n}$, and define $O_{i,n} \colon M_{i,n} \times M_{-i} \to \Delta(Z)$ by

$$\begin{aligned} &O_{i,n}(m_{i,n}, m_{-i})(z) \\ &= \alpha \Bigg(O_{i,n}^0(m_{i,n}, m_{-i,1}, \ldots, m_{-i,n-1})(z) \\ &\quad + \sum_{j \in I \setminus \{i\}} \sum_{k=n+1}^{N} \delta^{k-n} O_{j,k}^0(m_{j,k}, m_{-j,1}, \ldots, m_{-j,k-1})(z) \Bigg), \end{aligned}$$

where $m_{i,k} = m_{i,k}^*$ for $k \neq n$ when they appear in the second term, and

$$\alpha = 1 \Big/ \left(1 + (|I| - 1) \sum_{k=n+1}^{N} \delta^{k-n} \right)$$

is a normalization constant. Let $\Omega = \prod_{k=n}^{N} M_{-i,k}$. Since we chose sufficiently small δ, we have $\| O_{i,n}(\cdot, \cdot, \omega) - O_{i,n}^0 \| \leq \varepsilon_0 \leq \varepsilon_{n-1}$ for every $\omega \in \Omega$. Let $\nu_i^* \in \Delta(M_{-i} \times U_i \times H_{-i,n-1})$ be such that

$$\nu_i^*(E) = \nu_i(\{(m_{-i}, u_i, t_{-i}) \mid (m_{-i}, u_i, (\hat{\mu}_{j,k}(t_j))_{j \neq i, 1 \leq k \leq n-1}) \in E\})$$

for each measurable $E \subseteq M_{-i} \times U_i \times H_{-i,n-1}$. By the induction hypothesis,

$$\nu_i^* \left(\left\{ (m_{-i}, u_i, t_{-i,1}, \ldots, t_{-i,n-1}) \,\middle|\, \max_{j \neq i, 1 \leq k \leq n-1} d_{j,k}(t_{j,k}, m_{j,k}) \leq \varepsilon_{n-1} \right\} \right) = 1.$$

We also have $\mathrm{mrg}_{U_i \times H_{-i,n-1}} \nu_i^* = \hat{\mu}_{i,n}(t_i)$. Thus, we have $\mathrm{mrg}_{M_{-i} \times U_i} \nu_i^* \in \Delta_{\varepsilon_{n-1}, \hat{\mu}_{i,n}(t_i)}(M_{-i} \times U_i)$. Since $M_{i,n}$ is ε_{n-1}-dense in $\Delta(U_i \times H_{-i,n-1})$, there exists $m_{i,n}' \in M_{i,n}$ such that $d_{i,n}(\hat{\mu}_{i,n}(t_i), m_{i,n}') \leq \varepsilon_{n-1}$. By Lemma 1, $\mathrm{mrg}_{M_{-i} \times U_i} \nu_i^*$ strictly prefers $O_{i,n}(m_{i,n}', \cdot)$ to $O_{i,n}(m_{i,n}^*, \cdot)$, and thus $\mathrm{mrg}_{M_{-i} \times U_i} \nu_i^*$ strictly prefers $O(m_{i,n}', m_{i,-n}^*, \cdot)$ to $O^*(m_i^*, \cdot)$. This is a contradiction.

A.3. Proof of Proposition 3

Fix any equilibrium $\widehat{\sigma} = (\widehat{\sigma}_i)_{i \in I}$ of Bayesian game $(\widehat{\mathcal{T}}, \mathcal{M})$. Let $\sigma = (\sigma_i)_{i \in I}$ be a profile of mappings $\sigma_i \colon T_i \to \Delta(M_i)$ given by $\sigma_i = \widehat{\sigma}_i \circ \hat{\mu}_i$. For every $i \in I$, since $\widehat{\sigma}_i$ and $\hat{\mu}_i$ are both measurable, σ_i is also measurable. Also, for every $i \in I$ and every $m_i, m_i' \in M_i$ with $\sigma_i(t_i)(m_i) > 0$, we have

$$
\begin{aligned}
&\int_{U_i \times T_{-i}} \sum_{m_{-i} \in M_{-i}} \sum_{z \in Z} u_i(z)(O(m_i, m_{-i})(z) \\
&\quad - O(m_i', m_{-i})(z)) \left(\prod_{j \neq i} \sigma_j(t_j)(m_j) \right) \mu_i(t_i)(du_i, dt_{-i}) \\
&= \int_{U_i \times T_{-i}} \sum_{m_{-i} \in M_{-i}} \sum_{z \in Z} u_i(z)(O(m_i, m_{-i})(z) \\
&\quad - O(m_i', m_{-i})(z)) \left(\prod_{j \neq i} \widehat{\sigma}_j(\hat{\mu}_j(t_j))(m_j) \right) \mu_i(t_i)(du_i, dt_{-i}) \\
&= \int_{U_i \times \hat{T}_{-i}} \sum_{m_{-i} \in M_{-i}} \sum_{z \in Z} u_i(z)(O(m_i, m_{-i})(z) \\
&\quad - O(m_i', m_{-i})(z)) \left(\prod_{j \neq i} \widehat{\sigma}_j(\hat{t}_j)(m_j) \right) \mu_i^*(\hat{\mu}_i(t_i))(du_i, d\hat{t}_{-i}) \\
&\geq 0,
\end{aligned}
$$

where the second equality follows since $(\hat{\mu}_i)_{i \in I}$ is belief-preserving, and the last inequality follows since $\widehat{\sigma}_i(\hat{\mu}_i(t_i))(m_i) = \sigma_i(t_i)(m_i) > 0$ and $\widehat{\sigma}$ is an equilibrium of $(\widehat{\mathcal{T}}, \mathcal{M})$. Therefore, σ is an equilibrium of Bayesian game $(\mathcal{T}, \mathcal{M})$.

Appendix B. Interdependent preferences and λ-continuity

We present a formal and self-contained treatment of general interdependent expected utility preferences and the λ-continuity restriction. This treatment will be used in Appendix C.

One way to define state-dependent expected utility preferences for a general measurable space X is to have a preference $\succsim$ over acts over X represented by a belief about states $\mu \in \Delta(X)$ and a μ-integrable state-dependent utility index $u \colon X \times Z \to \mathbb{R}$ as follows:

$$
f \succsim f' \Leftrightarrow \int_X \sum_{z \in Z} u(x, z)(f(x)(z) - f'(x)(z))\mu(dx) \geq 0.
$$

Instead, we use a finite signed measure over $X \times Z$ to represent $\succsim$ as

$$
f \succsim f' \Leftrightarrow \int_{X \times Z} (f(x)(z) - f'(x)(z))\nu(dx, dz) \geq 0.
$$

The representation by a finite signed measure ν is formally equivalent to, via the Radon–Nikodym theorem, but more convenient than the representation by a belief-utility pair (μ, u). For example, u is meaningful only up to μ-null events, and hence multiple belief-utility pairs can represent the same preference. Indeed, although multiple signed measures can also represent the same preference, it is not difficult to pick a particular normalization. For example, if $\succsim$ is not completely indifferent over all outcomes, then we can choose $\overline{z}, \underline{z} \in Z$ such that $\overline{z} \succ \underline{z}$ and represent $\succsim$ uniquely by a signed measure ν over $X \times Z$ such that $\nu(X \times \{\overline{z}\}) = 1$ and $\nu(E \times \{\underline{z}\}) = 0$ for every measurable $E \subseteq X$.

In what follows, we use state-dependent expected utility preferences, described above, to define type spaces of interdependent preferences, interdependent preference hierarchies, and the universal type space. Along the way, we introduce various notions directly based on preferences so that we can guarantee easily that these notions are well defined and independent of representations and normalizations. But we also rephrase these notions, whenever possible, in terms of signed-measure representations to ease the reader into possibly unfamiliar notations.

Our exercise here is largely guided by the analogy between subjective beliefs and preferences, originated by Savage (1954) in single-agent environments and extended by Epstein and Wang (1996), Di Tillio (2008) and Ganguli et al. (2016) to multi-agent environments. At a technical level, our argument relies on mathematical similarities between probability measures and signed measures. At some subtle level, however, we need to understand a "patchwork" of possibly multiple signed-measure representations of a single preference, which we will discuss further in Appendix B.3.

B.1. State-dependent expected utility preferences

For a measurable space X, let $\mathrm{ca}(X)$ be the set of all finite signed measures over X. For $\nu \in \mathrm{ca}(X)$, $\|\nu\| = \sup_{\text{measurable } E,E' \subseteq X} (\nu(E) - \nu(E')) < \infty$ denotes the total variation of ν; $|\nu|$ denotes the total variation measure on X, defined by $|\nu|(E) = \|\nu(\cdot \cap E)\|$ for each measurable $E \subseteq X$. If X is a compact metric space, $\mathrm{ca}(X)$ is the dual of the set of continuous functions with the sup norm (the Riesz representation theorem).

Recall that $F(X)$ denotes the set of all acts over X, i.e., all measurable functions $f\colon X \to \Delta(Z)$. If X is a compact metric space, $F_c(X) \subseteq F(X)$ denotes the set of all continuous acts over X.

Let $P(X)$ be the set of all state-dependent expected utility preferences over $F(X)$ represented by $\nu \in \mathrm{ca}(X \times Z)$ as follows:

$$f \succsim f' \Leftrightarrow \int_{X \times Z} (f(x)(z) - f'(x)(z))\nu(dx, dz) \geq 0.$$

We say that a preference $\succsim \in P(X)$ is *certain of measurable* $E \subseteq X$ if $X \setminus E$ is Savage-null with respect to $\succsim$. For a preference $\succsim \in P(X)$ represented by $\nu \in \mathrm{ca}(X \times Z)$, $\succsim$ is certain of measurable $E \subseteq X$ if and only if $\nu(E' \times \{z\}) = \nu(E' \times \{z'\})$ for every measurable $E' \subseteq X \setminus E$ and every $z, z' \in Z$.

We endow $P(X)$ with the σ-algebra generated by $\{\succsim \in P(X) \mid f \succsim f'\}$ for each $f, f' \in F(X)$. If X is a compact metric space, we also endow $P(X)$ with the topology generated by $\{\succsim \in P(X) \mid f \succ f'\}$ for each $f, f' \in F_c(X)$; in this case, the Borel σ-algebra on $P(X)$ coincides with the original σ-algebra on $P(X)$.[28]

Given two measurable spaces X and Y, a measurable mapping $\varphi\colon X \to Y$ and a preference $\succsim \in P(X)$, we can define the *induced preference* $\varphi^P(\succsim)$ as the preference over $F(Y)$ such that,

[28] Since $F_c(X) \subseteq F(X)$, every Borel-measurable subset of $P(X)$ is measurable. Conversely, let $\mathcal{D} = \{E \subseteq X \mid E$ is Borel-measurable in X, and $\{\succsim \in P(X) \mid y_E y' \succsim y''_E y'''\}$ is Borel-measurable in $P(X)$ for every $y, y', y'', y''' \in \Delta(Z)\}$, where $y_E y'$ denotes the act over X that takes values y on E and y' on $X \setminus E$. Then $\mathcal{D}$ is a Dynkin system, and contains all closed subsets of X by Urysohn's lemma. Since the family of all closed subsets of X is a π-system, by the π-λ theorem, $\mathcal{D}$ coincides with the Borel σ-algebra on X. Hence $\{\succsim \in P(X) \mid f \succsim f'\}$ is Borel-measurable in $P(X)$ for all acts f and f' in the form of $y_E y'$ with Borel-measurable $E \subseteq X$. This extends to all simple acts and to all acts in the usual way.

for every $f, f' \in F(Y)$, it weakly prefers f to f' if and only if $\succsim$ weakly prefers $f \circ \varphi$ to $f' \circ \varphi$. (Note that $f \circ \varphi, f' \circ \varphi \in F(X)$.) It is easy to show that if $\succsim \in P(X)$ is represented by a signed measure $\nu \in \mathrm{ca}(X \times Z)$, then the induced preference $\varphi^P(\succsim)$ is represented by the induced signed measure $\nu' \in \mathrm{ca}(Y \times Z)$, where $\nu'(E) = \nu(\{(x, z) \in X \times Z \mid (\varphi(x), z) \in E\})$ for each measurable $E \subseteq Y \times Z$. We thus have $\varphi^P(\succsim) \in P(Y)$. Note that $\varphi^P \colon P(X) \to P(Y)$ is measurable; moreover, if X and Y are compact metric spaces and $\varphi \colon X \to Y$ is continuous, then $\varphi^P \colon P(X) \to P(Y)$ is also continuous.

The "marginal" is an important example of induced preferences. Given a product measurable space $X \times Y$ and a preference $\succsim \in P(X \times Y)$, the projection mapping $\mathrm{pr}_X \colon X \times Y \to X$ induces the *marginal*, denoted by $\mathrm{mrg}_X := (\mathrm{pr}_X)^P \colon P(X \times Y) \to P(X)$. In other words, given that we identify $F(X)$ as a subset of $F(X \times Y)$, where outcomes do not depend on the Y-coordinate, we define the marginal of $\succsim \in P(X \times Y)$ on X, $\mathrm{mrg}_X \succsim \in P(X)$, as the restriction of $\succsim$ to $F(X)$. This notion corresponds to the notion of marginal of a probability or signed measure. Indeed, if $\succsim$ is represented by a signed measure $\nu \in \mathrm{ca}(X \times Y \times Z)$, then $\mathrm{mrg}_X \succsim$ is represented by the marginal of ν on $X \times Z$, $\mathrm{mrg}_{X \times Z} \nu \in \mathrm{ca}(X \times Z)$, where $(\mathrm{mrg}_{X \times Z} \nu)(E \times \{z\}) = \nu(E \times Y \times \{z\})$ for each measurable $E \subseteq X$ and each $z \in Z$.

For a more specific example, consider a measurable space X, an arbitrary singleton set $\{*\}$ and a preference $\succsim \in P(X)$. Then the constant mapping from X to $\{*\}$ induces the restriction of the preference $\succsim$ to lotteries. If $\nu \in \mathrm{ca}(X \times Z)$ represents $\succsim$, then $\mathrm{mrg}_Z \nu \in \mathrm{ca}(Z) \cong \mathbb{R}^Z$ is a von Neumann–Morgenstern utility index that represents the restriction of the preference $\succsim$ to lotteries.

B.2. Type spaces and the universal type space

A *type space* is given by $\mathcal{T} = (T_i, \pi_i)_{i \in I}$, where, for each $i \in I$, T_i is a measurable space of agent i's types, and $\pi_i \colon T_i \to P(T_{-i})$ is a measurable mapping that maps his types to preferences.

Let $H_0 = \{*\}$ (an arbitrary singleton set) and $H_n = H_{n-1} \times P(H_{n-1}^{|I|-1}) = \prod_{k=0}^{n-1} P(H_k^{|I|-1})$ for each $n \geq 1$. Let $H = \prod_{n=0}^{\infty} P(H_n^{|I|-1})$ be the set of all hierarchies of interdependent preferences.

Given a type space $\mathcal{T} = (T_i, \pi_i)_{i \in I}$, we define the *interdependent preference hierarchy* of a type $t_i \in T_i$, $\hat{\pi}_i(t_i) = (\hat{\pi}_{i,1}(t_i), \hat{\pi}_{i,2}(t_i), \ldots)$, as follows: $\hat{\pi}_{i,1}(t_i)$ is the restriction of the preference $\pi_i(t_i)$ to lotteries, and for each $n \geq 2$, $\hat{\pi}_{i,n}(t_i)$ is the preference of type t_i over acts over the opponents' first $(n-1)$ order preferences, i.e., $\hat{\pi}_{i,n}(t_i) = (\hat{\pi}_{-i,1}, \ldots, \hat{\pi}_{-i,n-1})^P(\pi_i(t_i))$.[29] It is

[29] Recall the notion of induced preferences. For each $t_i \in T_i$, we define $\hat{\pi}_{i,1}(t_i) \in P(\{*\})$ by

$$\hat{\pi}_{i,1}(t_i) \text{ weakly prefers } y \text{ to } y' \Leftrightarrow \pi_i(t_i) \text{ weakly prefers } y \text{ to } y'$$

for each $y, y' \in F(\{*\}) = \Delta(Z)$.

For each $n \geq 2$ and each $f \in F(H_{n-1}^{|I|-1})$, we have $f \circ (\hat{\pi}_{-i,1}, \ldots, \hat{\pi}_{-i,n-1}) \in F(T_{-i})$ defined by

$$(f \circ (\hat{\pi}_{-i,1}, \ldots, \hat{\pi}_{-i,n-1}))(t_{-i}) = f((\hat{\pi}_{j,k}(t_j))_{j \neq i, 1 \leq k \leq n-1})$$

for each $t_{-i} \in T_{-i}$. Thus, for each $t_i \in T_i$, we define $\hat{\pi}_{i,n}(t_i) \in P(H_{n-1}^{|I|-1})$ by

$$\hat{\pi}_{i,n}(t_i) \text{ weakly prefers } f \text{ to } f' \Leftrightarrow \pi_i(t_i) \text{ weakly prefers } f \circ (\hat{\pi}_{-i,1}, \ldots, \hat{\pi}_{-i,n-1}) \text{ to } f' \circ (\hat{\pi}_{-i,1}, \ldots, \hat{\pi}_{-i,n-1})$$

for each $f, f' \in F(H_{n-1}^{|I|-1})$.

easy to show inductively that $\hat{\pi}_{i,n}: T_i \to P(H_{n-1}^{|I|-1})$ is measurable for every $n \geq 1$, and hence $\hat{\pi}_i: T_i \to \prod_{n=0}^{\infty} P(H_n^{|I|-1})$ is also measurable.

Following Heifetz and Samet (1998), we define T_i^* as the set of all interdependent preference hierarchies $t_i^* \in H$ such that $t_i^* = \hat{\pi}_i(t_i)$ for some type space $\mathcal{T} = (T_i, \pi_i)_{i \in I}$ and some type $t_i \in T_i$. We define $\pi_i^*: T_i^* \to P(T_{-i}^*)$ by

$$\pi_i^*(t_i^*) = \hat{\pi}_{-i}^P(\pi_i(t_i)),$$

where $t_i \in T_i$ in some type space $\mathcal{T} = (T_i, \pi_i)_{i \in I}$ such that $t_i^* = \hat{\pi}_i(t_i)$.[30] We can show that π_i^* is well defined (i.e., independent of the particular type space $\mathcal{T}$ and particular type t_i) and measurable.[31] We thus have the *universal type space* $\mathcal{T}^* = (T_i^*, \pi_i^*)_{i \in I}$. By construction, the profile $(\hat{\pi}_i)_{i \in I}$ of mappings $\hat{\pi}_i: T_i \to T_i^*$ is a preference-preserving morphism, also known as a type morphism in Heifetz and Samet (1998), from $\mathcal{T}$ to $\mathcal{T}^*$ in the following sense.[32]

Proposition 6. *For every type space $\mathcal{T} = (T_i, \pi_i)_{i \in I}$ and every agent $i \in I$, we have*

$$\pi_i^* \circ \hat{\pi}_i = \hat{\pi}_{-i}^P \circ \pi_i,$$

i.e., for every type $t_i \in T_i$ and every $f, f' \in F(T_{-i}^*)$, $\pi_i^*(\hat{\pi}_i(t_i))$ weakly prefers f to f' if and only if $\pi_i(t_i)$ weakly prefers $f \circ \hat{\pi}_{-i}$ to $f' \circ \hat{\pi}_{-i}$.

B.3. Compactness and metrizability of $P_\lambda(X)$

Let $P_0(X)$ be the set of preferences in $P(X)$ that are not completely indifferent over all outcomes. By excluding the preference that is completely indifferent over $F(X)$, we can show that $P_0(X)$ is Hausdorff if X is a compact metric space.[33]

Lemma 3. *If X is a compact metric space, then $P_0(X)$ is Hausdorff.*

---

[30] For each $f \in F(T_{-i}^*)$, we have $f \circ \hat{\pi}_{-i} \in F(T_{-i})$ defined by

$$(f \circ \hat{\pi}_{-i})(t_{-i}) = f((\hat{\pi}_j(t_j))_{j \neq i})$$

for each $t_{-i} \in T_{-i}$. Then we have $\pi_i^*(t_i^*) \in P(T_{-i}^*)$ defined by

$$\pi_i^*(t_i^*) \text{ weakly prefers } f \text{ to } f' \Leftrightarrow \pi_i(t_i) \text{ weakly prefers } f \circ \hat{\pi}_{-i} \text{ to } f' \circ \hat{\pi}_{-i}$$

for each $f, f' \in F(T_{-i}^*)$.

[31] For each $n \geq 0$, let $\mathrm{pr}_{-i,n}: T_{-i}^* \to H_n^{|I|-1}$ be the projection mapping. Fix any $f, f' \in F(H_n^{|I|-1})$. For each $t_i^* = (\succsim_1, \succsim_2, \ldots) \in T_i^*$, there exist a type space $\mathcal{T} = (T_i, \pi_i)_{i \in I}$ and a type $t_i \in T_i$ such that $t_i^* = \hat{\pi}_i(t_i)$. Then we have

$$\pi_i^*(t_i^*) \text{ weakly prefers } f \circ \mathrm{pr}_{-i,n} \text{ to } f' \circ \mathrm{pr}_{-i,n} \Leftrightarrow \pi_i(t_i) \text{ weakly prefers } f \circ \hat{\pi}_{-i,n}^P \text{ to } f' \circ \hat{\pi}_{-i,n}^P \Leftrightarrow \succsim_{n+1} \text{ weakly prefers } f \text{ to } f'.$$

Thus, $\{t_i^* \in T_i^* \mid \pi_i^*(t_i^*) \text{ weakly prefers } f \circ \mathrm{pr}_{-i,n} \text{ to } f' \circ \mathrm{pr}_{-i,n}\}$ is well defined and measurable. Since this is true for every n and every $f, f' \in F(H_n^{|I|-1})$, $\{t_i^* \in T_i^* \mid \pi_i^*(t_i^*) \text{ weakly prefers } f \text{ to } f'\}$ is well defined and measurable for every $f, f' \in F(T_{-i}^*)$, and hence $\pi_i^*: T_i^* \to P(T_{-i}^*)$ is well defined and measurable.

[32] In passing, we note that every preference-preserving morphism preserves interdependent preference hierarchies, and that $(\hat{\pi}_i)_{i \in I}$ is the unique preference-preserving morphism from $\mathcal{T}$ to $\mathcal{T}^*$.

[33] Indeed, Lemma 3 holds as long as we exclude the preference that is completely indifferent over $F(X)$.

Proof. Pick any pair of preferences $\succsim, \succsim' \in P_0(X)$ such that $\succsim \neq \succsim'$. Then there exist $f, f' \in F(X)$ such that $\succsim$ and $\succsim'$ have different preferences between f and f'. Since neither $\succsim$ nor $\succsim'$ is completely indifferent, we can assume without loss of generality that $f \succ f'$ and $f' \succ' f$.[34]

Let $\nu, \nu' \in \text{ca}(X \times Z)$ be finite signed measures that represent $\succsim$ and $\succsim'$, respectively. Applying Lusin's theorem to $(X \times Z, |\nu| + |\nu'|)$, we can assume without loss of generality that $f, f' \in F_c(X)$. Thus, $\succsim$ and $\succsim'$ are separated by two disjoint open sets generated by f and f'. □

We define λ-continuity as follows.

Definition 1. For a given $\lambda \in (0, 1/2]$, we say that a preference $\succsim$ is λ-*continuous* if there exist $\overline{z}, \underline{z} \in Z$ such that $\overline{z} \succ \underline{z}$ and $(1-\lambda)\,\overline{z} + \lambda f \succsim (1-\lambda)\,\underline{z} + \lambda f'$ for every $f, f' \in F(X)$.

If X is a compact metric space, then by Lusin's theorem, we can require $(1-\lambda)\,\overline{z} + \lambda f \succsim (1-\lambda)\,\underline{z} + \lambda f'$ only for all $f, f' \in F_c(X)$ without loss of generality.

Let $P_{\overline{z},\underline{z},\lambda}(X)$ be the set of all λ-continuous preferences for a fixed pair of outcomes $\overline{z}, \underline{z} \in Z$. Let $P_\lambda(X) = \bigcup_{\overline{z},\underline{z} \in Z} P_{\overline{z},\underline{z},\lambda}(X)$ be the set of all λ-continuous preferences.

Note that λ-continuity is preserved for induced preferences. That is, given each measurable mapping $\varphi\colon X \to Y$, if we have $\succsim \in P_{\overline{z},\underline{z},\lambda}(X)$, then we also have $\varphi^P(\succsim) \in P_{\overline{z},\underline{z},\lambda}(Y)$; if we have $\succsim \in P_\lambda(X)$, then we also have $\varphi^P(\succsim) \in P_\lambda(Y)$.

Fix a pair of outcomes $\overline{z}, \underline{z} \in Z$ and $\lambda \in (0, 1/2]$. Then each $\succsim \in P_{\overline{z},\underline{z},\lambda}(X)$ is uniquely represented by $\nu \in \text{ca}_{\overline{z},\underline{z},\lambda}(X \times Z)$, where

$$\text{ca}_{\overline{z},\underline{z},\lambda}(X \times Z) = \left\{ \nu \in \text{ca}(X \times Z) \;\middle|\; \begin{array}{l} \nu(X \times \{\overline{z}\}) = 1 \\ \nu(E \times \{\underline{z}\}) = 0 \text{ for every measurable } E \subseteq X \\ \int_{X \times Z} (f(x)(z) - f'(x)(z))\nu(dx, dz) \leq (1-\lambda)/\lambda \\ \text{for every } f, f' \in F(X) \end{array} \right\}.$$

In words, we normalize a signed-measure representation by first shifting the conditional expected utility of getting $\underline{z}$ given each event to 0, and then scaling the expected utility of getting $\overline{z}$ to 1. The condition that $\int_{X \times Z} (f(x)(z) - f'(x)(z))\nu(dx, dz) \leq (1-\lambda)/\lambda$ for every $f, f' \in F(X)$ is a rewriting of the definition of λ-continuity in terms of signed-measure representations. Via this normalization, $P_{\overline{z},\underline{z},\lambda}(X)$ is measurably isomorphic to $\text{ca}_{\overline{z},\underline{z},\lambda}(X \times Z)$; furthermore, if X is a compact metric space, then $P_{\overline{z},\underline{z},\lambda}(X)$ is topologically isomorphic (i.e., homeomorphic) to $\text{ca}_{\overline{z},\underline{z},\lambda}(X \times Z)$ endowed with the weak-* topology.

Note that this normalization is preserved for induced preferences. That is, given each measurable mapping $\varphi\colon X \to Y$, if ν belongs to $\text{ca}_{\overline{z},\underline{z},\lambda}(X \times Z)$, then the induced signed measure $\nu \circ (\varphi^{-1}, \text{id}_Z)$ belongs to $\text{ca}_{\overline{z},\underline{z},\lambda}(Y \times Z)$ with the same $\overline{z}, \underline{z} \in Z$ and $\lambda \in (0, 1/2]$. Therefore, if $\succsim \in P_{\overline{z},\underline{z},\lambda}(X)$ is represented by a normalized signed measure $\nu \in \text{ca}_{\overline{z},\underline{z},\lambda}(X \times Z)$, then the induced preference $\varphi^P(\succsim) \in P_{\overline{z},\underline{z},\lambda}(Y)$ is represented by the already normalized signed measure $\nu \circ (\varphi^{-1}, \text{id}_Z) \in \text{ca}_{\overline{z},\underline{z},\lambda}(Y \times Z)$.

Since each measurable function $g\colon X \times (Z \setminus \{\underline{z}\}) \to \mathbb{R}$ with $\|g\| \leq 1/(|Z| - 1)$ in the sup norm can be written as $g(x, z) = f(x)(z) - f'(x)(z)$ with some $f, f' \in F(X)$, we have $\|\nu\| \leq (|Z| - 1)(1-\lambda)/\lambda$ for every $\nu \in \text{ca}_{\overline{z},\underline{z},\lambda}(X \times Z)$. Conversely, since $\|f - f'\| \leq 1$ in the sup

[34] For example, if $f \sim f'$ and $f' \succ' f$, then pick $f'', f''' \in F(X)$ such that $f'' \succ f'''$. Then by slightly mixing f with f'' and f' with f''', we can make the first preference relation strict while maintaining the second preference relation.

norm for every $f, f' \in F(X)$, we have $\nu \in \mathrm{ca}^{\overline{z},\underline{z},\lambda}(X \times Z)$ for every $\nu \in \mathrm{ca}(X \times Z)$ such that $\nu(X \times \{\overline{z}\}) = 1$, $\nu(E \times \{\underline{z}\}) = 0$ for every measurable $E \subseteq X$, and $\|\nu\| \leq (1-\lambda)/\lambda$.

Lemma 4. *If X is a compact metric space, then $P^{\overline{z},\underline{z},\lambda}(X)$ is compact and metrizable for every $\overline{z}, \underline{z} \in Z$ and every $\lambda \in (0, 1/2]$.*

Proof. By the remark after Definition 1, $P^{\overline{z},\underline{z},\lambda}(X)$ is closed in $P_0(X)$. Also, $\mathrm{ca}^{\overline{z},\underline{z},\lambda}(X \times Z)$ can be seen as a subset of the ball $\{\nu \in \mathrm{ca}(X \times (Z \setminus \{\underline{z}\})) \mid \|\nu\| \leq (|Z|-1)(1-\lambda)/\lambda\}$, which is weak-* compact by the Riesz representation theorem and Alaoglu's theorem, and weak-* metrizable by the Stone–Weierstrass theorem. Thus, $\mathrm{ca}^{\overline{z},\underline{z},\lambda}(X \times Z)$ is compact and metrizable, and so is $P^{\overline{z},\underline{z},\lambda}(X)$. □

Note that Lemma 4 relies on λ-continuity with $\lambda \in (0, 1/2]$.

Recall that $P_\lambda(X) = \bigcup_{\overline{z},\underline{z} \in Z} P^{\overline{z},\underline{z},\lambda}(X)$. If $|Z| \geq 3$, then this union is not disjoint, i.e., a given preference $\succsim \in P_\lambda(X)$ may belong to $P^{\overline{z},\underline{z},\lambda}(X)$ with multiple pairs of $(\overline{z}, \underline{z})$. In this case, we do not choose any specific pair as "canonical". Instead, we view $P_\lambda(X)$ as a "patchwork" of finitely many $P^{\overline{z},\underline{z},\lambda}(X)$, each of which is homeomorphic to $\mathrm{ca}^{\overline{z},\underline{z},\lambda}(X \times Z)$.

Proposition 7. *If X is a compact metric space, then $P_\lambda(X)$ is compact and metrizable for every $\lambda \in (0, 1/2]$.*

Proof. By Lemmas 3 and 4, $P_\lambda(X)$ is a finite union of compact and metrizable subspaces $P^{\overline{z},\underline{z},\lambda}(X)$, and hence $P_\lambda(X)$ is compact and metrizable. (The metrizability follows from the Nagata–Smirnov metrization theorem. See Nagata (1985, Theorem 6.12).) □

B.4. λ-continuous type spaces

Fix $\lambda \in (0, 1/2]$. We say that a type space $\mathcal{T} = (T_i, \pi_i)_{i \in I}$ is *λ-continuous* if $\pi_i(t_i) \in P_\lambda(T_{-i})$ for every $i \in I$ and every $t_i \in T_i$. Note that in a λ-continuous type space, each type has a λ-continuous preference and common certainty of λ-continuous preferences. Moreover, the value of λ is fixed uniformly in types.

Let $H_{\lambda,0} = \{*\}$, $H_{\lambda,n} = H_{\lambda,n-1} \times P_\lambda(H_{\lambda,n-1}^{|I|-1})$ for each $n \geq 1$, and $H_\lambda = \prod_{n=0}^{\infty} P_\lambda(H_{\lambda,n}^{|I|-1})$. By Proposition 7, $H_{\lambda,n}$ is compact and metrizable for every $n \geq 0$. We endow H_λ with the product topology, and hence H_λ is also compact and metrizable. Since λ-continuity is preserved for induced preferences, the interdependent preference hierarchy of every λ-continuous type is also λ-continuous. That is, for every λ-continuous type space $\mathcal{T} = (T_i, \pi_i)_{i \in I}$ and every type $t_i \in T_i$, we have $\hat{\pi}_i(t_i) \in H_\lambda$. (Recall that $\hat{\pi}_i(t_i)$ denotes the interdependent preference hierarchy of t_i.)[35]

[35] Following Mertens and Zamir (1985) and Brandenburger and Dekel (1993), but replacing Kolmogorov's extension theorem by a version generalized to signed measures with uniformly bounded total variations, we can define the *universal λ-continuous type space* $\mathcal{T}^*_\lambda = (T^*_{i,\lambda}, \pi^*_{i,\lambda})_{i \in I}$ with the compact and metrizable set $T^*_{i,\lambda}$ of all λ-continuous preference hierarchies satisfying coherence and common certainty of coherence and the homeomorphism $\pi^*_{i,\lambda} = \pi^*_i|_{T^*_{i,\lambda}} : T^*_{i,\lambda} \to P_\lambda(T^*_{-i,\lambda})$. We do not need these facts, though.

Appendix C. Proofs in Section 5

The proofs in this section will use the self-contained treatment of general interdependent expected utility preferences in Appendix B.

C.1. Proof of Proposition 4

The following result generalizes Proposition 3.

Proposition 8. *For every type space $\mathcal{T} = (T_i, \pi_i)_{i \in I}$, every agent $i \in I$, and every type $t_i \in T_i$, we have*

$$E_i(t_i; \mathcal{T}, \mathcal{M}) \supseteq E_i(\hat{\pi}_i(t_i); \widehat{\mathcal{T}}, \mathcal{M})$$

for every finite mechanism $\mathcal{M} = ((M_i)_{i \in I}, O)$, where $\widehat{\mathcal{T}} = (\widehat{T}_i, \pi_i^*|_{\widehat{T}_i})_{i \in I}$ is a preference closed subspace of the universal type space $\mathcal{T}^* = (T_i^*, \pi_i^*)_{i \in I}$ with $\widehat{T}_i = \hat{\pi}_i(T_i)$ for each $i \in I$.

Proof. The proof is analogous to that of Proposition 3; we only need to replace the belief-preserving property by the preference-preserving property established in Proposition 6. □

Proposition 4 follows from Proposition 8 and the existence of equilibria for countable types (recall the proof of Theorem 2).

C.2. Proof of Proposition 5

Given a type space $\mathcal{T} = (T_i, \pi_i)_{i \in I}$ and a finite mechanism $\mathcal{M} = ((M_i)_{i \in I}, O)$, we define the set of actions that are *preference rationalizable for type t_i*, denoted by $PR_i(t_i)$ or $PR_i(t_i; \mathcal{T}, \mathcal{M})$ as follows:

$$PR_i^0(t_i) = M_i,$$

$$PR_i^{n+1}(t_i) = \left\{ m_i \in M_i \;\middle|\; \begin{array}{l} \text{there exists } \succsim_i \in P(M_{-i} \times T_{-i}) \text{ s.t.} \\ \text{(i) } \succsim_i \text{ is certain of } \prod_{j \neq i} \text{graph}(PR_j^n), \\ \text{(ii) } \text{mrg}_{T_{-i}} \succsim_i = \pi_i(t_i), \\ \text{(iii) } \succsim_i \text{ weakly prefers } O(m_i, \cdot) \text{ to } O(m_i', \cdot) \\ \quad \text{for every } m_i' \in M_i \end{array} \right\},$$

$$PR_i(t_i) = \bigcap_{n=0}^{\infty} PR_i^n(t_i).$$

Note that the inductive step is well defined since we can show inductively that $\text{graph}(PR_i^n)$ is measurable in $M_i \times T_i$ for every $i \in I$ and $n \geq 0$.

Lemma 5. *For every finite type space $\mathcal{T} = (T_i, \pi_i)_{i \in I}$ and every finite mechanism $\mathcal{M} = ((M_i)_{i \in I}, O)$, we have the following:*

1. *We have $m_i \notin PR_i^1(t_i)$ if and only if there exists $\sigma_i \in \Delta(M_i)$ such that:*
   (a) *$O(\sigma_i, m_{-i}) - O(m_i, m_{-i})$ is independent of $m_{-i} \in M_{-i}$, and*

(b) $\pi_i(t_i)$ *strictly prefers* $O(\sigma_i, m_{-i})$ *to* $O(m_i, m_{-i})$ *for some (and hence for all)* $m_{-i} \in M_{-i}$.[36]

2. $PR_i(t_i) = PR_i^1(t_i)$.

Proof. For part 1, the if direction is immediate. To show the only-if direction, let $\pi_i(t_i)$ be represented by $\bar{w}_i : T_{-i} \times Z \to \mathbb{R}$ as follows:

$$f \succsim f' \Leftrightarrow \sum_{t_{-i},z} (f(t_{-i})(z) - f'(t_{-i})(z))\bar{w}_i(t_{-i}, z) \geq 0.$$

If $m_i \notin PR_i^1(t_i)$, then there is no $w_i : M_{-i} \times T_{-i} \times Z \to \mathbb{R}$ such that

$$\sum_{m_{-i}} w_i(m_{-i}, t_{-i}, z) = \bar{w}_i(t_{-i}, z) \qquad \text{for all } t_{-i}, z,$$

$$\sum_{m_{-i},t_{-i},z} (O(m_i, m_{-i})(z) - O(m_i', m_{-i})(z))w_i(m_{-i}, t_{-i}, z) \geq 0 \qquad \text{for all } m_i'.$$

By Farkas' lemma, there exist $D : T_{-i} \times Z \to \mathbb{R}$ and $\sigma_i \in \Delta(M_i)$ such that

$$D(t_{-i}, z) - (O(\sigma_i, m_{-i})(z) - O(m_i, m_{-i})(z)) = 0 \qquad \text{for all } t_{-i}, m_{-i}, z,$$

$$\sum_{t_{-i},z} D(t_{-i}, z)\bar{w}_i(t_{-i}, z) > 0.$$

Thus, $O(\sigma_i, m_{-i})(z) - O(m_i, m_{-i})(z)$ is independent of m_{-i}, and $\pi_i(t_i)$ strictly prefers $O(\sigma_i, m_{-i})$ to $O(m_i, m_{-i})$.

For part 2, fix any player $i \in I$. For each $j \neq i$ and $t_j \in T_j$, if $m_j \in PR_j^1(t_j)$, then let $\sigma_j(m_j, t_j)$ be the point mass on m_j. If $m_j \notin PR_j^1(t_j)$, then by part 1, there exists $\sigma_j(m_j, t_j) \in \Delta(M_j)$ such that for every $z \in Z$, $O(\sigma_j(\cdot \mid m_j, t_j), m_{-j})(z) - O(m_j, m_{-j})(z)$ is independent of m_{-j}. Without loss of generality, we assume that $\sigma_j(m_j, t_j) \in \Delta(PR_j^1(t_j))$. For each $m_{-i} \in M_{-i}$ and $t_{-i} \in T_{-i}$, define $\sigma_{-i}(m_{-i}, t_{-i}) \in \Delta(\prod_{j\neq i} PR_j^1(t_j))$ by $\sigma_{-i}(m_{-i}, t_{-i})(m'_{-i}) = \prod_{j\neq i} \sigma_j(m_j, t_j)(m'_j)$ for each $m'_{-i} \in PR_{-i}^1(t_{-i})$.

Pick any $t_i \in T_i$ and any $m_i \in PR_i^1(t_i)$. Then there exists $\succsim_i \in P(M_{-i} \times T_{-i})$ such that $\mathrm{mrg}_{T_{-i}} \succsim_i = \pi_i(t_i)$ and m_i is a best response with respect to $\succsim_i$. We will show that m_i survives in the second step of iteration.

Let $\pi_i(t_i)$ be represented by $\bar{w}_i : T_{-i} \times Z \to \mathbb{R}$. Let $\succsim_i$ be represented by $w_i : M_{-i} \times T_{-i} \times Z \to \mathbb{R}$ such that $\sum_{m_{-i}} w_i(m_{-i}, \cdot, \cdot) = \bar{w}_i$. Define $w_i' : M_{-i} \times T_{-i} \times Z \to \mathbb{R}$ by

$$w_i'(m'_{-i}, t_{-i}, z) = \sum_{m_{-i}} \sigma_{-i}(m_{-i}, t_{-i})(m'_{-i})w_i(m_{-i}, t_{-i}, z)$$

for $m'_{-i} \in M_{-i}$, $t_{-i} \in T_{-i}$ and $z \in Z$. Denote by $\succsim_i' \in P(M_{-i} \times T_{-i})$ the preference represented by w_i'. First, since $\sigma_{-i}(m_{-i}, t_{-i}) \in \Delta(\prod_{j\neq i} PR_j^1(t_j))$ for every $m_{-i} \in M_{-i}$ and every $t_{-i} \in T_{-i}$, $\succsim_i'$ is certain of $\prod_{j\neq i} \mathrm{graph}(PR_j^1)$. Second, we have

[36] We define $O(\sigma_i, m_{-i})$ by $O(\sigma_i, m_{-i})(z) = \sum_{m_i'} O(m_i', m_{-i})(z)\sigma_i(m_i')$ for each $z \in Z$.

$$\sum_{m'_{-i}} w'_i(m'_{-i}, t_{-i}, z) = \sum_{m_{-i}, m'_{-i}} \sigma_{-i}(m_{-i}, t_{-i})(m'_{-i}) w_i(m_{-i}, t_{-i}, z)$$
$$= \sum_{m_{-i}} w_i(m_{-i}, t_{-i}, z)$$
$$= \bar{w}_i(t_{-i}, z),$$

and hence $\text{mrg}_{T_{-i}} \succsim'_i = \pi_i(t_i)$. Third, for every $m'_i \in M_i$, we have

$$\sum_{m'_{-i}} O(m'_i, m'_{-i})(z) w'_i(m'_{-i}, t_{-i}, z)$$
$$= \sum_{m_{-i}, m'_{-i}} O(m'_i, m'_{-i})(z) \sigma_{-i}(m_{-i}, t_{-i})(m'_{-i}) w_i(m_{-i}, t_{-i}, z)$$
$$= \sum_{m_{-i}} O(m'_i, \sigma_{-i}(m_{-i}, t_{-i}))(z) w_i(m_{-i}, t_{-i}, z)$$
$$= \sum_{m_{-i}} (O(m'_i, m_{-i})(z) + D(m_{-i}, t_{-i}, z)) w_i(m_{-i}, t_{-i}, z)$$
$$= \sum_{m_{-i}} O(m'_i, m_{-i})(z) w_i(m_{-i}, t_{-i}, z) + \sum_{m_{-i}} D(m_{-i}, t_{-i}, z) w_i(m_{-i}, t_{-i}, z)$$

for every $t_{-i} \in T_{-i}$ and every $z \in Z$, where $D(m_{-i}, t_{-i}, z) := O(m'_i, \sigma_{-i}(m_{-i}, t_{-i}))(z) - O(m'_i, m_{-i})(z)$ is independent of m'_i by the construction of $\sigma_{-i}(m_{-i}, t_{-i})$. Since m_i is a best response with respect to $\succsim_i$ represented by w_i, it is also a best response with respect to $\succsim'_i$ represented by w'_i. □

Lemma 6. *For every pair of finite type spaces $\mathcal{T} = (T_i, \pi_i)_{i \in I}$ and $\mathcal{T}' = (T'_i, \pi'_i)_{i \in I}$, every agent $i \in I$, and every pair of types $t_i \in T_i$ and $t'_i \in T'_i$, if $\hat{\pi}_{i,1}(t_i; \mathcal{T}) = \hat{\pi}_{i,1}(t'_i; \mathcal{T}')$, then we have*

$$PR_i(t_i; \mathcal{T}, \mathcal{M}) = PR_i(t'_i; \mathcal{T}', \mathcal{M})$$

for every finite mechanism $\mathcal{M} = ((M_i)_{i \in I}, O)$.

Proof. The result follows from Lemma 5. □

Proposition 5 follows from rewriting the statement of Lemma 6 in terms of equilibrium.

C.3. The robust scoring rule

As in Section 4.2, we analyze a single-agent mechanism that reveals her state-dependent preferences. Fix $\lambda \in (0, 1/2]$. Fix a compact metric space X with metric d. By Proposition 7, $P_\lambda(X)$ is also a compact metric space, whose metric is denoted by d_P. The choice function with respect to $\succsim \in P_\lambda(X)$ is given by

$$C_{\succsim}(f, f') = \begin{cases} f & \text{if } \succsim \text{ weakly prefers } f \text{ to } f', \\ f' & \text{if } \succsim \text{ strictly prefers } f' \text{ to } f \end{cases}$$

for every $f, f' \in F(X)$.

By the Stone–Weierstrass theorem, there exists a countable dense subset $F = \{f_1, f_2, \ldots\} \subset F_c(X)$ in the sup norm.

We consider the following direct mechanism $\mathcal{M}^0 = (M^0, O^0)$ for a single agent with message set $M^0 = P_\lambda(X)$ and outcome function $O^0 \colon M^0 \times X \to \Delta(Z)$ given by

$$O^0(m, x)(z) = \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} 2^{-k-l} C_m(f_k, f_l)(x)(z) \tag{3}$$

for each realized state $x \in X$ and reported preference $m \in M^0$.

For each $\delta > 0$, $\succsim \in P_\lambda(X)$, and measurable space Ω, we define

$$P_{\lambda,\delta,\succsim}(X \times \Omega) = \left\{ \succsim'' \in P_\lambda(X \times \Omega) \;\middle|\; \begin{array}{l} \text{there exists } \succsim' \in P_\lambda(X \times X' \times \Omega) \\ \quad \text{with } X' = X \text{ s.t.} \\ \text{(i) } \succsim' \text{ is certain of } \{(x, x') \mid d(x, x') \le \delta\} \times \Omega, \\ \text{(ii) } \mathrm{mrg}_X \succsim' = \succsim, \\ \text{(iii) } \mathrm{mrg}_{X' \times \Omega} \succsim' = \succsim'' \end{array} \right\}.$$

Lemma 7. *Fix $\lambda \in (0, 1/2]$. For every $\varepsilon > 0$, there exists $\delta > 0$ such that the following is true for every preference $\succsim \in P_\lambda(X)$, every pair of messages m, m', every measurable space Ω, and every perturbed outcome function $O \colon M^0 \times X \times \Omega \to \Delta(Z)$: if $d_P(\succsim, m) \le \delta$, $d_\Delta(\succsim, m') > \varepsilon$, and $\|O(\cdot, \cdot, \omega) - O^0\| \le \delta$ for every $\omega \in \Omega$, then every preference in $P_{\lambda,\delta,\succsim}(X \times \Omega)$ strictly prefers $O(m, \cdot, \cdot)$ to $O(m', \cdot, \cdot)$.*

Proof. Suppose not. Then there exists $\varepsilon > 0$ such that for every $n \in \mathbb{N}$, there exist $\succsim_n, m_n, m'_n \in P_\lambda(X)$ with $d_P(\succsim_n, m_n) \le 1/n$ and $d_P(\succsim_n, m'_n) > \varepsilon$, measurable space Ω_n, perturbed outcome function $O_n \colon M^0 \times X \times \Omega_n \to \Delta(Z)$ with $\|O_n(\cdot, \cdot, \omega) - O^0\| \le 1/n$ for every $\omega \in \Omega_n$, $\succsim'_n \in P_\lambda(X \times X' \times \Omega_n)$ with $X' = X$ such that $\succsim'_n$ is certain of $\{(x, x') \mid d(x, x') \le \delta\} \times \Omega_n$, $\mathrm{mrg}_X \succsim'_n = \succsim_n$, and $\mathrm{mrg}_{X' \times \Omega} \succsim'_n$ weakly prefers $O_n(m'_n, \cdot, \cdot)$ to $O_n(m_n, \cdot, \cdot)$. By taking a subsequence if necessary, we can assume without loss of generality that $\succsim'_n \in P_{\overline{z}, \underline{z}, \lambda}(X \times X' \times \Omega_n)$ with a fixed pair $(\overline{z}, \underline{z})$, and hence $\succsim_n \in P_{\overline{z}, \underline{z}, \lambda}(X)$ with the same pair $(\overline{z}, \underline{z})$. By Proposition 7, by taking a subsequence if necessary, we can find $\succsim^*, m'^* \in P_\lambda(X)$ such that $\succsim_n \to \succsim^*$ and $m'_n \to m'^*$ as $n \to \infty$. Note that $m_n \to \succsim^*$ as $n \to \infty$, and $\succsim^* \neq m'^*$. Also note that $\succsim^* \in P_{\overline{z}, \underline{z}, \lambda}(X)$. Let $\nu_n, \nu^* \in \mathrm{ca}_{\overline{z}, \underline{z}, \lambda}(X \times Z)$ and $\nu'_n \in \mathrm{ca}_{\overline{z}, \underline{z}, \lambda}(X \times X' \times \Omega_n \times Z)$ represent $\succsim_n$, $\succsim^*$, and $\succsim'_n$, respectively. Note that $\mathrm{mrg}_{1,4} \nu'_n = \nu_n$.

Let

$$u^* = \int O^0(\succsim^*, x)(z) d\nu^*(x, z).$$

Claim 3. *We have*

$$\lim_{n \to \infty} \int O^0(m_n, x)(z) d\nu_n(x, z) = u^*,$$

$$\limsup_{n \to \infty} \int O^0(m'_n, x)(z) d\nu_n(x, z) < u^*.$$

Proof of Claim 3. The claim follows from showing that

$$\lim_{n\to\infty}\int C_{m_n}(f_k,f_l)(x)(z)d\nu_n(x,z)=\int C_{\succsim^*}(f_k,f_l)(x)(z)d\nu^*(x,z),$$
$$\limsup_{n\to\infty}\int C_{m'_n}(f_k,f_l)(x)(z)d\nu_n(x,z)\le\int C_{\succsim^*}(f_k,f_l)(x)(z)d\nu^*(x,z),$$

for each k,l, and that the second inequality holds with strict inequality for some k,l. The first equality and the second weak inequality follow from the standard revealed preference argument. To show the strict inequality, since $\succsim^*\neq m'^{,*}$ and $F\subset F_c(X)$ is dense in the sup norm, there exist k,l such that $\succsim^*$ strictly prefers f_k to f_l while $m'^{,*}$ strictly prefers f_l to f_k. Since m'_n strictly prefers f_l to f_k for sufficiently large n, we have

$$\begin{aligned}
&\lim_{n\to\infty}\int C_{m'_n}(f_k,f_l)(x)(z)d\nu_n(x,z)\\
&=\lim_{n\to\infty}\int f_l(x)(z)d\nu_n(x,z)\\
&=\int f_l(x)(z)d\nu^*(x,z)\\
&<\int f_k(x)(z)d\nu^*(x,z)\\
&=\int C_{\succsim^*}(f_k,f_l)(x)(z)d\nu^*(x,z).\quad\square
\end{aligned}$$

Claim 4. *We have*

$$\lim_{n\to\infty}\left(\int O_n(m,x',\omega)(z)d\nu'_n(x,x',\omega,z)-\int O^0(m,x)(z)d\nu_n(x,z)\right)=0$$

and the convergence is uniform in $m\in M^0$.

Proof of Claim 4. Note that

$$\begin{aligned}
&\left|\int O_n(m,x',\omega)(z)d\nu'_n(x,x',\omega,z)-\int O^0(m,x)(z)d\nu_n(x,z)\right|\\
&\le\left|\int O_n(m,x',\omega)(z)d\nu'_n(x,x',\omega,z)-\int O^0(m,x')(z)d\nu'_n(x,x',\omega,z)\right|\\
&\quad+\left|\int O^0(m,x')(z)d\nu'_n(x,x',\omega,z)-\int O^0(m,x)(z)d\nu_n(x,z)\right|.
\end{aligned}$$

The first term is bounded above by $\sup_{\omega\in\Omega_n}\|O_n(\cdot,\cdot,\omega)-O^0\|\|\nu'_n\|\le(|Z|-1)(1-\lambda)/(n\lambda)$.

To show that the second term converges to 0 uniformly in m, it is enough to show that

$$\lim_{n\to\infty}\left(\int f(x')(z)d\nu'_n(x,x',\omega,z)-\int f(x)(z)d\nu_n(x,z)\right)=0$$

for each $f\in F_c(X)$. Since X is a compact metric space, f is uniformly continuous. Therefore, for every $\eta>0$, there exists N such that $\max_z\left|f(x)(z)-f(x')(z)\right|<\eta$ whenever $d(x,x')\le 1/N$. For every $n\ge N$, we have

$$\left|\int f(x')(z)d\nu'_n(x,x',\omega,z) - \int f(x)(z)d\nu_n(x,z)\right|$$
$$\leq \left|\int f(x')(z)d\nu'_n(x,x',\omega,z) - \int f(x)(z)d\nu'_n(x,x',\omega,z)\right|$$
$$+ \left|\int f(x)(z)d\nu'_n(x,x',\omega,z) - \int f(x)(z)d\nu_n(x,z)\right|.$$

The first term is bounded above by $\eta\|\nu'_n\| \leq \eta(|Z|-1)(1-\lambda)/\lambda$; the second term is equal to zero since $\text{mrg}_{X\times Z}\nu'_n = \nu_n$. □

Claims 3 and 4 contradict the assumption that $\text{mrg}_{X'\times\Omega}\succsim'_n$ weakly prefers $O_n(m'_n,\cdot,\cdot)$ to $O_n(m_n,\cdot,\cdot)$. □

C.4. Proof of Theorem 4

Fix $\lambda \in (0,1/2]$. Given a λ-continuous type space $\mathcal{T} = (T_i,\pi_i)_{i\in I}$ and a finite mechanism $\mathcal{M} = ((M_i)_{i\in I}, O)$, we define the set of actions that are λ-*continuously rationalizable for type* t_i, denoted by $R_{i,\lambda}(t_i)$ or $R_{i,\lambda}(t_i;\mathcal{T},\mathcal{M})$, as follows:

$$R^0_{i,\lambda}(t_i) = M_i,$$
$$R^{n+1}_{i,\lambda}(t_i) = \left\{ m_i \in M_i \;\middle|\; \begin{array}{l} \text{there exists } \succsim_i \in P_\lambda(M_{-i}\times T_{-i}) \text{ s.t.} \\ \text{(i) } \succsim_i \text{ is certain of } \prod_{j\neq i}\text{graph}(R^n_{j,\lambda}), \\ \text{(ii) } \text{mrg}_{T_{-i}}\succsim_i = \pi_i(t_i), \\ \text{(iii) } \succsim_i \text{ weakly prefers } O(m_i,\cdot) \text{ to } O(m'_i,\cdot) \\ \qquad \text{for every } m'_i \in M_i \end{array} \right\},$$
$$R_{i,\lambda}(t_i) = \bigcap_{n=0}^{\infty} R^n_{i,\lambda}(t_i).$$

Note that the inductive step is well defined since we can show inductively that $\text{graph}(R^n_{i,\lambda})$ is measurable in $M_i\times T_i$ for every $i\in I$ and $n\geq 0$.

Let d_λ be a metric compatible with the product topology on the set H_λ of λ-continuous preference hierarchies.

Proposition 9. *Fix $\lambda\in(0,1/2]$. For every $\varepsilon>0$, there exists a finite mechanism $\mathcal{M} = ((M_i)_{i\in I}, O)$ such that*

$$d_\lambda(\hat\pi_i(t_i;\mathcal{T}),\hat\pi_i(t'_i;\mathcal{T}')) > \varepsilon \Rightarrow R_{i,\lambda}(t_i;\mathcal{T},\mathcal{M}) \cap R_{i,\lambda}(t'_i;\mathcal{T}',\mathcal{M}) = \emptyset$$

for every pair of λ-continuous type spaces $\mathcal{T} = (T_i,\pi_i)_{i\in I}$ and $\mathcal{T}' = (T'_i,\pi'_i)_{i\in I}$, every agent $i\in I$, and every pair of types $t_i\in T_i$ and $t'_i\in T'_i$.

Sketch of the Proof. The proof is analogous to that of Proposition 2. By Proposition 7, $H^{|I|-1}_{\lambda,n-1}$ is compact and metrizable, and hence we can let $X = H^{|I|-1}_{\lambda,n-1}$ and apply Lemma 7 repeatedly. □

Proof of Theorem 4. $1\Rightarrow 3$ follows from Proposition 4. $3\Rightarrow 2$ follows from the fact that equilibrium is a refinement of λ-continuous preference rationalizability. $2\Rightarrow 1$, or its contrapositive $\neg 1\Rightarrow\neg 2$, follows from Proposition 9. □

References

Abreu, D., Matsushima, H., 1992. Virtual Implementation in Iteratively Undominated Strategies: Incomplete Information. Discussion paper. Princeton University, University of Tokyo.

Abreu, D., Sen, A., 1991. Virtual implementation in Nash equilibrium. Econometrica 59, 997–1021.

Afriat, S., 1967. The construction of utility functions from expenditure data. Int. Econ. Rev. 8, 67–77.

Battigalli, P., Dufwenberg, M., 2009. Dynamic psychological games. J. Econ. Theory 144, 1–35.

Bergemann, D., Morris, S., 2005. Robust mechanism design. Econometrica 73, 1771–1813.

Bergemann, D., Morris, S., 2009a. Robust implementation in direct mechanisms. Rev. Econ. Stud. 76, 1175–1206.

Bergemann, D., Morris, S., 2009b. Robust virtual implementation. Theor. Econ. 4, 45–88.

Bergemann, D., Morris, S., 2012. Robust Mechanism Design. World Scientific Publishing, Singapore.

Brandenburger, A., Dekel, E., 1993. Hierarchies of belief and common knowledge. J. Econ. Theory 59, 189–198.

Chambers, C., 2008. Proper scoring rules for general decision models. Games Econ. Behav. 63, 32–40.

Chambers, C., Lambert, N., 2014. Dynamically eliciting unobservable information. In: EC '14 Proceedings of the 15th ACM Conference on Economics and Computation, pp. 987–988.

Dasgupta, P., Maskin, E., 2000. Efficient auctions. Q. J. Econ. 115, 341–388.

Dekel, E., Fudenberg, D., Morris, S., 2006. Topologies on types. Theor. Econ. 1, 275–309.

Dekel, E., Fudenberg, D., Morris, S., 2007. Interim correlated rationalizability. Theor. Econ. 2, 15–40.

Di Tillio, A., 2008. Subjective expected utility in games. Theor. Econ. 3, 287–323.

Duggan, J., 1997. Virtual Bayesian implementation. Econometrica 65, 1175–1199.

Ely, J.C., Peski, M., 2006. Hierarchies of belief and interim rationalizability. Theor. Econ. 1, 19–65.

Epstein, L., Wang, T., 1996. 'Beliefs about beliefs' without probabilities. Econometrica 64, 1343–1373.

Friedenberg, A., Meier, M., 2015. The context of a game. Econ. Theory. http://dx.doi.org/10.1007/s00199-015-0938-z.

Ganguli, J., Heifetz, A., Lee, B., 2016. Universal interactive preferences. J. Econ. Theory 162, 237–260.

Geanakoplos, J., Pearce, D., Stacchetti, E., 1989. Psychological games and sequential rationality. Games Econ. Behav. 1, 60–79.

Gossner, O., Mertens, J., 2001. The Value of Information in Zero-Sum Games. Université Paris-Nanterre and CORE, Université Catholique de Louvain.

Grant, S., Meneghel, I., Tourky, R., 2016. Savage games. Theor. Econ. 11, 641–682.

Gul, F., Pesendorfer, W., 2016. Interdependent preference models as a theory of intentions. J. Econ. Theory 165, 179–208.

Heifetz, A., Samet, D., 1998. Topology-free typology of beliefs. J. Econ. Theory 82, 324–341.

Hellman, Z., 2014. A game with no Bayesian approximate equilibria. J. Econ. Theory 153, 138–151.

Jackson, M., 1991. Bayesian implementation. Econometrica 59, 461–477.

Levine, D., 1998. Modeling altruism and spitefulness in experiments. Rev. Econ. Dyn. 1, 593–622.

Mertens, J., Zamir, S., 1985. Formalization of Bayesian analysis for games with incomplete information. Int. J. Game Theory 14, 1–29.

Morris, S., Takahashi, S., 2012. Games in Preference Form and Preference Rationalizability. Discussion paper. Princeton University.

Nagata, J., 1985. Modern General Topology. North-Holland.

Palfrey, T., Srivastava, S., 1989. Mechanism design with incomplete information: a solution to the implementation problem. J. Polit. Econ. 97, 668–691.

Sadzik, T., 2010. Beliefs Revealed in Bayesian-Nash Equilibrium. Discussion paper. New York University.

Savage, L., 1954. The Foundations of Statistics, 1st edn. Wiley, New York.

Serrano, R., Vohra, R., 2010. Multiplicity of mixed equilibria in mechanisms: a unified approach to exact and approximate implementation. J. Math. Econ. 46, 775–785.

Simon, R., 2003. Games of incomplete information, ergodic theory, and the measurability of equilibria. Isr. J. Math. 138, 73–92.

Sprumont, Y., 2000. On the testable implications of collective choice theories. J. Econ. Theory 93, 205–232.

Yildiz, M., 2015. Invariance to representation of information. Games Econ. Behav. 94, 142–156.