Interdependent Preferences and Strategic Distinguishability[*]

Dirk Bergemann[†]   Stephen Morris[‡]   Satoru Takahashi[§]

September 2010

Abstract

A universal type space of interdependent expected utility preference types is constructed from higher-order preference hierarchies describing (i) an agent's (unconditional) preferences over a lottery space; (ii) the agent's preference over Anscombe-Aumann acts conditional on the unconditional preferences; and so on.

Two types are said to be strategically indistinguishable if they have an equilibrium action in common in any mechanism that they play. We show that two types are strategically indistinguishable if and only if they have the same preference hierarchy. We examine how this result extends to alternative solution concepts and strategic relations between types.

KEYWORDS: Interdependent Preferences, Higher-Order Preference Hierarchy, Universal Type Space, Strategic Distinguishability.

JEL CLASSIFICATION: C79, D82, D83.

[†]Yale University, dirk.bergemann@yale.edu
[‡]Princeton University, smorris@princeton.edu
[§]Princeton University, satorut@princeton.edu

1 Introduction

Economists often assume that agents' preferences are interdependent for informational or psychological reasons. We know how to use Harsanyi type spaces to represent many kinds of such interdependence of preferences. In this paper, we characterize when two types are strategically distinguishable in the sense that they are guaranteed to behave differently in some finite mechanism mapping actions to outcomes.

Our characterization uses a universal type space of interdependent, higher-order, preferences of a finite set of agents, analogous to the universal space of higher-order beliefs introduced by Mertens and Zamir (1985). We assume common certainty that (i) agents are expected utility maximizers; (ii) agents are not indifferent between all outcomes; and (iii) there is a worst outcome for each agent. The universal space is mathematically isomorphic to the Mertens-Zamir universal belief space (although it has a very different interpretation). We show that two types are strategically distinguishable if and only if they map to different points in the universal space of interdependent preferences.

This result gives a clean and straightforward answer to the question: what can you observe (and be certain to observe) about agents' interdependent preferences by seeing how they play games, i.e., behave in strategic environments? Our answer is:

1. You can learn an agent's first order (or unconditional) preferences: what are his preferences over outcomes unconditional on anything other agents do or say?

2. Since you can learn all agents' unconditional preferences, you can also learn an agent's second order preferences: what are his preferences over acts that are contingent on the first order preferences of other agents?

3. And then you can learn his third order preferences. And so on.

You cannot learn any more than this. This implies, in particular, that it is not possible to distinguish between informational and psychological reasons for interdependence. And it implies that interdependence of preferences can be observed only when there is uncertainty about preferences, i.e., when I expect my preference to change on observing your preferences.

There are (at least) a couple of reasons why we believe that a systematic study of strategic distinguishability may be of interest. First, economists' traditional view of preferences is that they are not directly observed but are best understood as being revealed by agents' choices in actual or hypothetical decision problems, and there is a developed revealed preference theory of individual

2

choice behavior; we see this paper as being a step towards a strategic revealed preference theory.[1] Second, the content of the specific modelling assumptions is not always transparent and this is especially true when talking about interdependent preferences. By mapping all types into a canonical universal interdependent type space, we provide a clear operational definition of interdependent types.

Our main result concerns one solution concept, equilibrium, and one equivalence class on agents' interdependent types, strategic indistinguishability. We also discuss what happens if we consider an appropriate but very permissive definition of rationalizability for our environment—dubbed interim preference correlated rationalizability (IPCR)—and an alternative, more refined, equivalence class on agents' types: two types are said to be strategically equivalent if they have the same set of rationalizable actions in all strategic environments (strategic distinguishability only required a non-empty intersection of those sets). We show that the same universal interdependent preference space characterizes strategic distinguishability for IPCR, and thus for any solution concept which refines IPCR and coarsens equilibrium. We also show that the universal interdependent preference space characterizes strategic equivalence for IPCR, so that, for IPCR, two types are strategically distinguishable if and only if they are strategically equivalent. But for equilibrium, more information than that contained in the universal interdependent preference space is required to capture strategic equivalence (as shown by an example in Section 3).

We maintain the worst outcome assumption in order to exclude trivial types that are completely indifferent over all outcomes and to maintain compactness of our type spaces which is necessary for our results. In Section 8.1, we discuss how the worst outcome assumption can be relaxed while maintaining non-triviality and compactness of preferences.

Our results are closely tied to a number of existing literatures. Most importantly, our formal contribution can be viewed as an extension of results of Abreu and Matsushima (1992) from finite to more general type spaces. They 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. A necessary condition is a "measurability" condition that, in the language of this paper, requires that the social choice function gives the same outcome to strategically indistinguishable types. They provide a characterization of the measurability condition that essentially states that types are strategically distinguishable if and only if they differ in their preference hierarchies. Iterated deletion of strictly dominated strategies is equivalent to a refined version of rationalizability—interim correlated rationalizability—that is intermediate between equilibrium and IPCR. They also show that the measurability condition is necessary for virtual Bayesian

---

[1]This is discussed further in Section 8.3.

implementation in equilibrium, and so their argument establishes a characterization of strategic distinguishability for equilibrium as well. We extend the analysis of Abreu and Matsushima (1992) to infinite type spaces. As well as raising new technical challenges, a benefit of the extension is that the equivalence relation between preference hierarchies and strategic distinguishability can be stated in terms of a universal space and thus without reference to a specific type space from which the types are drawn. In Section 8.2, we discuss how the analysis in this paper is related to Bergemann and Morris (2009), which showed that robust virtual implementation is possible only if there is not too much interdependence in preferences.

As we noted above, our universal interdependent preference space construction is mathematically equivalent to the construction of the universal belief space of Mertens and Zamir (1985), although we are giving it a quite different interpretation. Epstein and Wang (1996) construct a universal space of hierarchies of non-expected utility preferences, incorporating non-expected utility preferences such as ambiguity aversion, but maintaining monotonicity as well as additional regularity conditions. We must dispense with monotonicity to incorporate the interdependence of preferences we want to capture. We relax monotonicity to the worst outcome assumption, but impose independence to get an expected utility representation. Di Tillio (2008) allows general preferences, and thus does not require Epstein and Wang's monotonicity condition or independence, but restricts attention to preferences over finite outcomes at every level of the hierarchy.

A number of authors have considered problems that arise in behaviorally identifying psychologically motivated properties of preferences that involve interdependence (see Levine (1998) and Weibull (2004)) such as conditional altruism (e.g. I want to be generous only to those people who are generous themselves). Our leading example below will concern conditional altruism. Motivated by such problems, Gul and Pesendorfer (2007) construct a universal space of interdependent preference types and our objective of constructing a universal interdependent preference space follows their exercise. They identify a maximal set of types which captures all distinctions that can be expressed in a natural language. When they consider applications of their universal space to incomplete information settings, they treat incomplete information separately and thus they do not address the interaction (and indistinguishability in a state dependent expected utility setting) of beliefs and utilities. Our focus is on static games and solution concepts (equilibrium and rationalizability) without sequential rationality or other refinements of those solution concepts. This implies that, in a complete information setting, it is not possible to identify any interdependence in agents' types (a point emphasized in our leading example of Section 3). Thus our universal space of interdependent types ends up being much coarser than that of Gul and Pesendorfer (2007). In particular, their types reflect much counterfactual information (what preferences would be conditional on other agents' types) that cannot be strategically distinguished in our sense. An interesting topic

4

for future work is the extent to which finer behavioral distinctions, such as strategic equivalence, and dynamic games with sequential rationality refinements (where behavior will reflect counterfactual information) can reveal the fine information contained in Gul and Pesendorfer (2007) types. Recent work on dynamic mechanism design in "payoff type" environments, as Müller (2009) and Penta (2009), may be relevant for such extensions.

A recent literature (Dekel, Fudenberg and Morris [DFM] (2006, 2007), Ely and Pęski (2006), Liu (2009), Sadzik (2009)) has examined what can be learned about agents' beliefs and higher-order beliefs about a state space Θ when it is (informally) assumed that there is common certainty of agents' "payoffs" as a function of their actions in a game and the realized state $\theta \in \Theta$. Our results can be understood as a relaxation of the assumption of common certainty of payoffs in that literature. In particular, that literature can be summarized as follows. DFM show that two types have the same interim correlated rationalizable (ICR) actions if and only if they have the same higher-order beliefs, i.e., they map to the same Mertens and Zamir [MZ] (1985) type. Thus, in the language of this paper, MZ types characterize strategic equivalence for ICR under the common certainty of payoffs assumption. ICR is a permissive solution concept that allows agents' actions to reveal information about others' actions and the payoff relevant state. If restrictions are put on what can be revealed, as in the interim independent rationalizability (IIR) of DFM (2007), then finer distinctions over types are required to characterize strategic equivalence. Ely and Pęski (2006) describe richer hierarchies than MZ types which characterize IIR in two player games. Liu (2009) and Sadzik (2009) discuss even richer information needed to characterize Bayesian Nash equilibrium (BNE). Although not highlighted in this literature, it is easy to deduce from these existing results that MZ types characterize strategic indistinguishability for all three solution concepts (ICR, IIR and BNE); in other words, two types have an ICR/IIR/BNE action in common in every mechanism if and only if they have the same MZ type. To see why, note that we can always find a BNE action they have in common by looking for pooling equilibria where redundant information is ignored. Thus a summary of the "common certainty of payoffs" literature is:

	strategically equivalent	strategically indistinguishable
ICR	Mertens-Zamir space	Mertens-Zamir space
IIR	Ely-Pęski space	Mertens-Zamir space
BNE	richer Liu/Sadzik space	Mertens-Zamir space

Our results in this paper offer a clean generalization of this picture. This literature combines beliefs and higher-order beliefs about some payoff relevant states with common certainty of a mapping from action profiles and payoff relevant states to payoffs. Relaxing the common certainty of payoffs assumption, we must construct a universal space of higher-order (expected utility) preferences. We

show that this characterizes strategic indistinguishability for equilibrium, for IPCR and for any solution concept in between. We show that it also characterizes strategic equivalence for IPCR but not necessarily for more refined versions of rationalizability and equilibrium.

The paper is organized as follows. Section 2 describes our setting and poses the strategic distinguishability question for equilibrium. Section 3 considers in detail an example with conditional altruism to motivate the approach and results in the paper. Section 4 describes the construction of the universal space of interdependent preferences. Section 5 reports our main result: our universal space characterizes equilibrium strategic distinguishability. Section 6 introduces the solution concept of interim preference correlated rationalizability, and presents the proof that our universal space characterizes strategic distinguishability for equilibrium, IPCR and everything in between. Section 7 formally introduces the finer strategic equivalence relation, shows that our universal space characterizes IPCR strategic equivalence and discusses the formal connection with the common certainty of payoffs literature. Section 8 concludes.

2 The Setting and Benchmark Question

An outside observer will see a finite set of agents, $\mathcal{I} = \{1, ..., I\}$, making choices in strategic situations, where there is a finite set of outcomes Z and a compact and metrizable set of observable states Θ. We will maintain the assumption that, for each agent i, there is an outcome $w_i \in Z$ which is a worst outcome for that agent; in Section 8.1, we discuss relaxations of this assumption.

We are interested in what the outside observer can infer about agents' (perhaps interdependent) preferences by observing agents' rational choices in strategic situations. We will consider standard Harsanyi type space models of agents' perhaps interdependent preferences. A type space consists of a measurable set of unobservable states, Ω, and for each agent i, a measurable space of types T_i, a measurable belief function $\nu_i \colon T_i \to \Delta\left(\Theta \times \Omega \times T_{-i}\right)$ and a bounded and measurable utility function $u_i \colon \Theta \times \Omega \times T \times Z \to \mathbb{R}$. Consistent with the assumption that agent i has a worst outcome $w_i \in Z$, we require

$$u_i\left(\theta, \omega, t, z\right) \geq u_i\left(\theta, \omega, t, w_i\right)$$

for all $\theta \in \Theta$, $\omega \in \Omega$, $t \in T$ and $z \in Z$. In addition, we will make the non-triviality assumption that for every $t_i \in T_i$ and $\nu_i\left(\cdot \,|t_i\right)$-almost every $\left(\theta, \omega, t_{-i}\right) \in \Theta \times \Omega \times T_{-i}$, there exists some $z \in Z$ such that $u_i\left(\theta, \omega, t, z\right) > u_i\left(\theta, \omega, t, w_i\right)$. Thus a Harsanyi type space is given by $\mathcal{T} = \left(\Omega, \left(T_i, \nu_i, u_i\right)_{i \in \mathcal{I}}\right)$.

We define a belief-closed subset of the type space to be a product set of agents' types where each agent is sure to be in that subset. Formally, a product set $\tilde{T} = \prod_i \tilde{T}_i$ of types with measurable $\tilde{T}_i \subseteq T_i$ is belief-closed if for every $i \in \mathcal{I}$ and $t_i \in \tilde{T}_i$, $\nu_i\left(\Theta \times \Omega \times \tilde{T}_{-i} \,|t_i\right) = 1$.

A strategic situation is modelled as a mechanism, where each agent i has a finite set of actions A_i and an outcome function $g\colon \Theta \times A \to \Delta\left(Z\right)$. Thus a mechanism is defined by $\mathcal{M} = \left(\left(A_i\right)_{i\in\mathcal{I}}, g\right)$.

The pair $\left(\mathcal{T}, \mathcal{M}\right)$ describes a game of incomplete information. A strategy for agent i in this game is a measurable function $\sigma_i\colon T_i \to \Delta\left(A_i\right)$. We extend the domain of g to mixed strategies in the usual way. Bayesian Nash equilibria do not always exist on large type spaces. However, even when equilibria do not exist on large type spaces, equilibria may exist on belief-closed subsets of the large type space. We will follow Sadzik (2010) in defining such "local" equilibria.

Definition 1 *A strategy profile $\sigma = \left(\sigma_i\right)_{i\in\mathcal{I}}$ is a local equilibrium of the game $\left(\mathcal{T}, \mathcal{M}\right)$ on the belief-closed subspace $\tilde{T}$ if, for every $i \in \mathcal{I}$ and $t_i \in \tilde{T}_i$, $\sigma_i\left(t_i\right)$ maximizes*

$$\int_{\Theta\times\Omega\times T_{-i}} u_i(\theta, \omega, (t_i, t_{-i}), g\left(\theta, (a_i, \sigma_{-i}(t_{-i}))\right))d\nu_i(t_i)(\theta, \omega, t_{-i}).$$

Let $E_i(t_i, \mathcal{T}, \mathcal{M})$ be the set of all local equilibrium actions of type t_i, i.e., the set of actions played with positive probability by t_i in any local equilibrium of $\left(\mathcal{T}, \mathcal{M}\right)$ on any belief-closed subspace $\tilde{T}$ with $t_i \in \tilde{T}_i$.

We say that a type t_i is countable if there exists a countable belief-closed subspace $\tilde{T} = \prod_j \tilde{T}_j$ with $t_i \in \tilde{T}_i$. By Kakutani's fixed-point theorem, $E_i(t_i, \mathcal{T}, \mathcal{M}) \neq \emptyset$ if t_i is countable.

The main relation between types that we seek to characterize in this paper is the following.

Definition 2 *Two types of agent i, t_i in $\mathcal{T}$ and t_i' in $\mathcal{T}'$, are strategically indistinguishable if, for every mechanism $\mathcal{M}$, there exists some action that can be chosen by both types, so that*

$$E_i(t_i, \mathcal{T}, \mathcal{M}) \cap E_i(t_i', \mathcal{T}', \mathcal{M}) \neq \emptyset$$

for every $\mathcal{M}$. Conversely, t_i and t_i' are strategically distinguishable if there exists a mechanism in which no action can be chosen by both types, so that

$$E_i(t_i, \mathcal{T}, \mathcal{M}^*) \cap E_i(t_i', \mathcal{T}', \mathcal{M}^*) = \emptyset$$

for some $\mathcal{M}^*$.

Our main result will be a characterization of strategic distinguishability. Before reporting our result, we report examples to motivate and provide intuition for results.

3 Examples and Motivation

In this section, we illustrate by means of examples, that there are many equivalent ways of describing a type's beliefs and utilities, which all give rise to the same preferences and thus behavior. We refer to this multiplicity in the representation of preferences as redundancy. Our purpose in analyzing these examples is to describe the redundancy, use it to motivate a canonical representation of interdependent types, and give an intuition why this representation exactly captures strategic distinguishability as described in the previous section. We begin with the well known case of decision-theoretic redundancy and then discuss strategic redundancy.

3.1 Decision Theoretic Redundancy

Consider the following setting. There are two agents 1 and 2. There are three outcomes: each agent receives nothing (outcome 0); agent 1 receives a prize (outcome 1); or agent 2 receives the prize (outcome 2). Thus the environment consists of two agents, an outcome space $Z = \{0, 1, 2\}$ and no observable states. Outcome 0 will be a worst outcome for both agents.

The Harsanyi type space can be described as follows. There are two equally likely but unobservable states, L and H, interpreted as representing situations where the agent are either unrelated or siblings. Each agent i observes a conditionally independent signal $s_i \in \{l, h\}$, with $\Pr(l|L) = \Pr(h|H) = \frac{2}{3}$. If the agents are unrelated, then each one cares only about the probability that he/she gets the prize. If the agents are siblings, then each one is altruistic with weight $\frac{1}{2}$ on the sibling's consumption. Thus agent 1 is indifferent between a 50% chance of getting the prize for herself and agent 2 getting the prize. Now the type space consists of unobservable states $\Omega = \{L, H\}$; type spaces $T_1 = T_2 = \{l, h\}$; and utility functions of the form

$$
u_i(\omega, t, z) = u_i(\omega, (t_1, t_2), z) = \begin{cases} 0, \text{ if } z = 0 \\ 1, \text{ if } z = i \\ \frac{1}{2}, \text{ if } z = 3 - i \text{ and } \omega = H \\ 0, \text{ if } z = 3 - i \text{ and } \omega = L \end{cases} ;
$$

and beliefs on $\Omega \times T_1 \times T_2$ consistent with the following common prior:

$\omega = L:$

$t_1 \backslash t_2$	l	h
l	$\frac{2}{9}$	$\frac{1}{9}$
h	$\frac{1}{9}$	$\frac{1}{18}$

$\omega = H:$

$t_1 \backslash t_2$	l	h
l	$\frac{1}{18}$	$\frac{1}{9}$
h	$\frac{1}{9}$	$\frac{2}{9}$

In this example, there are no observable states Θ, and so we suppress them in our notation.

While this is one formal representation, there are many equivalent ways of describing a type's beliefs and utilities that give rise to the same preferences and thus behavior. We refer to this as

decision theoretic redundancy. A first simple and well known observation is that states that are not observed by any agent are redundant and can be integrated out. See, for example, Milgrom (2004), page 159, for a discussion of this point. Thus an alternative Harsanyi type space representation of the above example is the following. There are no unobservable states, the type spaces remain $T_1 = T_2 = \{l, h\}$; but the utility functions have the form:

$$u_i\left((t_1, t_2), z\right) = \begin{cases} 0, \text{ if } z = 0 \\ 1, \text{ if } z = i \\ \frac{1}{10}, \text{ if } z = 3 - i \text{ and } (t_1, t_2) = (l, l) \\ \frac{1}{4}, \text{ if } z = 3 - i \text{ and } (t_1, t_2) = (l, h) \text{ or } (h, l) \\ \frac{2}{5}, \text{ if } z = 3 - i \text{ and } (t_1, t_2) = (h, h) \end{cases} \quad ;$$

and beliefs on $T_1 \times T_2$ consistent with the following common prior:

$t_1 \backslash t_2$	l	h
l	$\frac{5}{18}$	$\frac{2}{9}$
h	$\frac{2}{9}$	$\frac{5}{18}$

This model has a natural interpretation in terms of conditional altruism: type h is nice and type l is nasty. Both types are altruistic, type h is more altruistic than type l, and both are more altruistic when the other is more altruistic. This mirrors the modelling of conditional altruism in, for example, Levine (1998). But importantly, there is no way to distinguish conditional altruism (intuitively, a property of the "utility") from the informational story of a sibling relationship.

Another redundancy in the description of Harsanyi types is that since the utility function is allowed to depend on types, the distinction between "utility" and "beliefs" is arbitrary and all we can observe is the product of the two. Another way of making this point is to observe that the choice of numeraire is arbitrary but affects whether independence is reflected in beliefs or utilities. We can illustrate this with three more equivalent representations of the above example.

One re-normalized representation is to let an agent assign utility 1 to the other agent getting the prize, and adjust beliefs and other utilities accordingly. This gives rise to the following utility functions:

$$u_i\left((t_1, t_2), z\right) = \begin{cases} 0, \text{ if } z = 0 \\ 10, \text{ if } z = i \text{ and } (t_1, t_2) = (l, l) \\ 4, \text{ if } z = i \text{ and } (t_1, t_2) = (l, h) \text{ or } (h, l) \\ \frac{5}{2}, \text{ if } z = i \text{ and } (t_1, t_2) = (h, h) \\ 1, \text{ if } z = 3 - i \end{cases} \quad ;$$

where the beliefs on $T_1 \times T_2$ must now be adjusted to ensure that the product of probabilities and

9

an agent's utility of an outcome remain in the same proportion:

$t_1 \backslash t_2$	l	h
l	$\frac{1}{9}$	$\frac{2}{9}$
h	$\frac{2}{9}$	$\frac{4}{9}$

A second re-normalization is to let the numeraire take the canonical form of a uniform distribution over all non-worst outcomes. In this case, we normalize the expected utility of a 50/50 lottery between the siblings getting a prize equal to 1. This gives rise to the utility functions:

$$u_i\left((t_1,t_2),z\right) = \begin{cases} 0, \text{ if } z = 0 \\ \frac{20}{11}, \text{ if } z = i \text{ and } (t_1,t_2) = (l,l) \\ \frac{8}{5}, \text{ if } z = i \text{ and } (t_1,t_2) = (l,h) \text{ or } (h,l) \\ \frac{10}{7}, \text{ if } z = i \text{ and } (t_1,t_2) = (h,h) \\ \frac{2}{11}, \text{ if } z = 3-i \text{ and } (t_1,t_2) = (l,l) \\ \frac{2}{5}, \text{ if } z = 3-i \text{ and } (t_1,t_2) = (l,h) \text{ or } (h,l) \\ \frac{4}{7}, \text{ if } z = 3-i \text{ and } (t_1,t_2) = (h,h) \end{cases} ;$$

and beliefs on $T_1 \times T_2$ consistent with the common prior:

$t_1 \backslash t_2$	l	h
l	$\frac{11}{45}$	$\frac{2}{9}$
h	$\frac{2}{9}$	$\frac{14}{45}$

The correlation of types changed as we changed the numeraire. In fact, we can choose beliefs any way we like and still find an equivalent representation.

A third re-normalization comes from letting beliefs be uniform and independent:

$t_1 \backslash t_2$	l	h
l	$\frac{1}{4}$	$\frac{1}{4}$
h	$\frac{1}{4}$	$\frac{1}{4}$

,

implying that the utility functions must then be proportional to:

$$u_i\left((t_1,t_2),z\right) = \begin{cases} 0, \text{ if } z = 0 \\ 10, \text{ if } z = i \text{ and } (t_1,t_2) = (l,l) \text{ or } (h,h) \\ 8, \text{ if } z = i \text{ and } (t_1,t_2) = (l,h) \text{ or } (h,l) \\ 1, \text{ if } z = 3-i \text{ and } (t_1,t_2) = (l,l) \\ 2, \text{ if } z = 3-i \text{ and } (t_1,t_2) = (l,h) \text{ or } (h,l) \\ 4, \text{ if } z = 3-i \text{ and } (t_1,t_2) = (h,h) \end{cases} .$$

10

These re-normalizations are possible because of the well-known property of state dependent utility representations of expected utility preferences that they do not pin down probabilities. This point is discussed, for example, in Myerson (1991) (on page 72) - he labels two incomplete information games where one is such a re-normalization of the other as representing "fully equivalent games" - and it is well known to empirical researchers on auctions (see Paarsch and Hong (2006)). He labels two incomplete information games where one is such a re-normalization of the other as representing "fully equivalent games".

Our solution to these two forms of decision theoretic redundancy (unobserved states and inseparability of beliefs and utilities) will be to work with preference type spaces, where unobserved states are integrated and types are identified with preferences over Anscombe-Aumann acts contingent on observable states and others' types. Thus we will abstract from numeraires, beliefs and utilities in the preference type space representation. In the example, type h of agent 1 will be characterized by the fact that if given a choice between $f \colon T_2 \to \Delta(Z)$ and $f' \colon T_2 \to \Delta(Z)$, he will weakly prefer f to f' if and only if

$$
\begin{aligned}
& 5f(h)(1) + 4f(l)(1) + 2f(h)(2) + f(l)(2) \\
\geq \ & 5f'(h)(1) + 4f'(l)(1) + 2f'(h)(2) + f'(l)(2)
\end{aligned}
$$

Mapping Harsanyi type spaces into preference type spaces is straightforward. However, it neither gives a natural language to express types nor ensures that types are strategically distinguishable. We will therefore introduce a natural canonical way to represent interdependent types. Consider type h of agent 1. What can we say about how this type will behave in different strategic situations? A first level observation is that this type will have an unconditional altruism (i.e., the marginal rate of substitution between her opponent and herself getting the prize) of $\frac{1}{3}$ ($= \frac{2}{3} \times \frac{1}{2}$). This is all we could find out about this type's preferences in a single person choice setting. But in a richer strategic setting, we could identify that type's altruism conditional on her opponent's unconditional altruism. In particular, type h of agent 1 will not give up anything in exchange for prizes conditional on agent 2's unconditional altruism being anything other than $\frac{1}{3}$ or $\frac{1}{6}$. Conditional on agent 2's unconditional altruism being $\frac{1}{6}$, agent 1 would be prepared to "pay" (i.e., give up unconditional probability of getting the prize) $\frac{4}{9}$ for the prize and $\frac{1}{9}$ ($= \frac{4}{9} \times \frac{1}{4}$) for agent 2 to get the prize. Conditional on agent 2's unconditional altruism being $\frac{1}{3}$, agent 1 would be prepared to pay $\frac{5}{9}$ for the prize and $\frac{2}{9}$ ($= \frac{5}{9} \times \frac{2}{5}$) for agent 2 to get the prize.

Our main result will involve a generalization of this description. In Section 4, we provide a formal description of a universal space of possible expected utility types, consisting of (i) unconditional (expected utility) preferences; (ii) preferences conditional on others' unconditional preferences; and so on. In Section 5, we confirm that two types are guaranteed to behave differently in equilibrium

of some mechanism if and only if they correspond to different types in this universal space. But before we move to the general analysis, we will give another example demonstrating how two types that may look quite different in a preference type space, and are decision theoretically distinct, map to the same preference hierarchy in the universal type space. We refer to this phenomenon as strategic redundancy.

3.2 Strategic Redundancy

Suppose we start with the example described earlier. But now assume that, in addition to agent i's altruism generated by kinship, she puts an additional weight on her sister's consumption of $\frac{1}{6}$ only if $s_i = l$. Now we will have the same common prior probability distribution over type profiles (t_1, t_2):

$t_1 \backslash t_2$	l	h
l	$\frac{5}{18}$	$\frac{2}{9}$
h	$\frac{2}{9}$	$\frac{5}{18}$

but now we add $\frac{1}{6}$ to the agent's type conditional on being the low type only, giving

$$
u_i \left((t_1, t_2), z \right) = \begin{cases} 0, \text{ if } z = 0 \\ 1, \text{ if } z = i \\ \frac{4}{15}, \text{ if } z_i = 3 - i \text{ and } (t_i, t_j) = (l, l) \\ \frac{1}{4}, \text{ if } z_i = 3 - i \text{ and } (t_i, t_j) = (h, l) \\ \frac{5}{12}, \text{ if } z_i = 3 - i \text{ and } (t_i, t_j) = (l, h) \\ \frac{2}{5}, \text{ if } z_i = 3 - i \text{ and } (t_i, t_j) = (h, h) \end{cases} ;
$$

Now the unconditional altruism of the low type of each agent is $\frac{5}{9} \left(\frac{4}{15} \right) + \frac{4}{9} \left(\frac{5}{12} \right) = \frac{1}{3}$, while the unconditional altruism of the high type of each agent is $\frac{5}{9} \left(\frac{2}{5} \right) + \frac{4}{9} \left(\frac{1}{4} \right) = \frac{1}{3}$, i.e., the same. This immediately implies that both types will map to the same type in the universal preference space, and will therefore be strategically indistinguishable from each other in equilibrium, and from any "complete information" type with common certainty that the unconditional altruism is $\frac{1}{3}$. This example illustrates a form of redundancy analogous to (but different from) the redundancy present in Mertens and Zamir (1985) and Dekel, Fudenberg and Morris (2007).

While types l, h and the complete information type with unconditional altruism $\frac{1}{3}$ are strategically indistinguishable, it is easy to construct a game where equilibrium actions of one type are not equilibrium actions of the other type. Suppose agent 1 has action U and D and agent 2 has actions L and R. The following table shows the probabilities that agents 1 and 2 get the prize as

a function of the action profile:

	L	R
U	$\frac{1}{2}, \frac{1}{2}$	$\frac{1}{3}, \frac{1}{3}$
D	$\frac{151}{240}, 0$	$\frac{1}{3} - \varepsilon, \frac{1}{3} - \varepsilon$

for some small $\varepsilon > 0$.

The payoffs in the game are asymmetric. Each agent has the same probability of getting the prize unless action profile (D, L) is chosen. Observe that if agent 1 chooses D rather than U when her opponent plays L, she is getting a $\frac{31}{120}$ probability of getting the prize for each unit of probability that agent 2's probability is reduced. She will only want to do this if her relative valuation of agent 2 getting the prize is less than $\frac{31}{120}$.

Now observe that on the "reduced" complete information type space (without redundant types), agent 1 must choose U in equilibrium: if she expects agent 2 to choose R, this gives strictly higher probability to both agents getting the prize; if she expects agent 2 to choose L, her valuation of agent 2 getting the prize is $\frac{1}{3}$, which is more than $\frac{31}{120}$.

On the "rich" Harsanyi type space (with redundant types), there will also be an equilibrium where all types of the two agents choose U and L respectively. Thus types with the same preference hierarchy do indeed have an equilibrium action in common, as shown by our main theorem. However, there will also a strict equilibrium where, for agent 1, type l chooses U and type h chooses D, and for agent 2, type l chooses L and type h chooses R. To see why, note that if agent 1 was type h and sure that agent 2 was type l choosing L, she would have a strict incentive to choose D, as her valuation of agent 2 getting the prize is $\frac{1}{4}$, which is less than $\frac{31}{120}$; and ε gain to choosing U if agent 2 chooses R will not reverse this strict preference for small ε. On the other hand, if agent 1 was type l and sure that agent 2 was type l choosing L she would have a strict incentive to choose U, as her valuation of agent 2 getting the prize is $\frac{4}{15}$, which is more than $\frac{31}{120}$; and ε gain to choosing U if agent 2 chooses R will ensure she wants to choose U.

Now observe that for agent 2, independent of her type, L is a best response if she expects agent 1 to be type l playing U, and R is a best response if she expects agent 1 to be type h playing D. But now one can verify that, as types are correlated and each type attaches probability $\frac{5}{9}$ to the other agent being the same type, agent 2 has a best response to play L if type l and R if type h. The detailed calculations appear in the Appendix A.

This example illustrates that while the complete information type and the rich type are guaranteed to have an equilibrium action in common, they may not have the same set of equilibrium actions. In this sense, the types are not strategically equivalent. This feature is not special to the equilibrium solution concept. U is also the unique undominated action, in the sense that it is the unique best response for any beliefs about agent 2's action given the complete information

preferences, and thus it is the unique interim correlated rationalizable (ICR) action (Dekel, Fudenberg and Morris (2007)) for the complete information type (by the same argument that worked for equilibrium), but since D is an equilibrium action of type l in the rich Harsanyi type space, it must also be an ICR action.

In section 6, we study both strategic indistinguishability and strategic equivalence under alternative solutions concepts, equilibrium and various versions of rationalizability. We will show that our strategic indistinguishability result holds for all the alternative solution concepts, as illustrated by this example.

But we will see that there are subtleties in defining rationalizable outcomes. In the solution concept of ICR, each agent is allowed to have conjectures in which opponents' actions and observable states are correlated in the minds of the agent, so that the opponents' actions reveal information about the observable state in the agent's mind. Analogously, in our context, it is natural to allow agents' conjectures to reveal information about their own preferences. We will formally describe a generalization of ICR, called interim preference correlated rationalizability, and show that two types are strategically equivalent under this solution concept if and only if they map to the same type in the universal space of interdependent preferences. However, for more refined solution concepts (such as equilibrium and interim correlated rationalizability), strategical equivalence generates a finer partition than our universal space.

We can illustrate this with our example. Consider the complete information type, with common certainty that each agent's unconditional altruism is $\frac{1}{3}$. But suppose that each agent's unconditional altruism arose from the belief that the other agent is a sibling with probability $\frac{2}{3}$ and not with probability $\frac{1}{3}$, and suppose that agent 1 believed that her opponent was going to choose R only if he was her sibling. Then D would be a best response. Thus both actions are interim preference correlated rationalizable in this example. As the example illustrates, this solution concept is extremely permissive.

4 Preference Types

We introduce preference type spaces that capture interdependent preferences and have no decision theoretic redundancy. We then construct a universal preference type space, which consists of preference hierarchies.

4.1 State-Dependent Preferences

We first define state-dependent preferences for a single agent in the framework of Anscombe and Aumann (1963). We begin with a measurable space X of states and a finite set Z of outcomes with

$|Z| \geq 2$. An (*Anscombe-Aumann*) *act* is a measurable mapping from X to $\Delta(Z)$. The set of all such acts is denoted by $F(X)$ and endowed with the sup norm. For $y, y' \in \Delta(Z)$ and measurable $E \subseteq X$, $y_E y'$ is the act that yields the lottery y over E and the lottery y' over $X \setminus E$. We consider the following conditions on binary relation $\succsim$ over $F(X)$. For a fixed worst outcome $w \in Z$, we define $P_w(X)$ to be the set of all binary relations over $F(X)$ that have a non-trivial state-dependent expected utility representation respecting the worst outcome:

Definition 3 *A binary relation over $F(X)$ is a (w worst outcome) expected utility preference if there exists $\mu \in \Delta(X \times (Z \setminus \{w\}))$ that satisfies*

$$f \succsim f' \Leftrightarrow \int_{X \times (Z \setminus \{w\})} f(x)(z) d\mu(x, z) \geq \int_{X \times (Z \setminus \{w\})} f'(x)(z) d\mu(x, z)$$

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

This representation can be axiomatized with a simple variant of standard arguments in decision theory.

1. *completeness*: for every $f, f' \in F(X)$, $f \succsim f'$ or $f' \succsim f$.

2. *transitivity*: for every $f, f', f'' \in F(X)$, if $f \succsim f'$ and $f' \succsim f''$, then $f \succsim f''$.

3. *independence*: for every $f, f', f'' \in F(X)$ and $\lambda \in (0, 1]$, $f \succsim f'$ if and only if $\lambda f + (1 - \lambda) f'' \succsim \lambda f' + (1 - \lambda) f''$.

4. *continuity*: for every $f, f', f'' \in F(X)$, if $f \succ f' \succ f''$, then there exists $\varepsilon \in (0, 1)$ such that $(1 - \varepsilon) f + \varepsilon f'' \succ f' \succ (1 - \varepsilon) f'' + \varepsilon f$.

5. *monotone continuity*: for every $z, z', z'' \in Z$ with $z \succ z'$ and decreasing sequence $\{E_n\}_{n \in \mathbb{N}}$ of measurable subsets of X with $\bigcap_n E_n = \emptyset$, there exists $n \in \mathbb{N}$ such that $z''_{E_n} z \succ z'$ and $z \succ z''_{E_n} z'$.

6. *non-triviality*: there exist $f, f' \in F(X)$ with $f \succ f'$.

7. *w worst outcome*: for every $f \in F(X)$, $f \succsim w$.

Proposition 1 $\succsim \in P_w(X)$ *if and only if it satisfies completeness, transitivity, independence, continuity, monotone continuity, non-triviality and w worst outcome.*

An event $E \subseteq X$ is $\succsim$-*null* if $z_E w \sim w$ for every $z \in Z$. For $\succsim$ represented by $\mu \in \Delta(X \times (Z \setminus \{w\}))$, E is $\succsim$-null if and only if $\mu(E \times (Z \setminus \{w\})) = 0$. An event E is $\succsim$-*certain* if $X \setminus E$ is $\succsim$-null.

For a preference $\succsim \in P_w(X)$ and a measurable space Y, a measurable mapping $\varphi \colon X \to Y$ induces a preference $\varphi^P(\succsim) \in P_w(Y)$ given by

$$f \; \varphi^P(\succsim) \; f' \Leftrightarrow f \circ \varphi \succsim f' \circ \varphi$$

for any $f, f' \in F(Y)$. In particular, for a preference $\succsim \in P_w(X \times Y)$, the projection from $X \times Y$ to X induces the *marginal preference of* $\succsim$, $\mathrm{mrg}_X \succsim \in P_w(X)$, which is the restriction of $\succsim$ to acts over $X \times Y$ that do not depend on the Y-coordinate.

$P_w(X)$ is treated as a measurable space with the σ-algebra generated by $\{\succsim \in P_w(X) \mid f \succsim f'\}$ for any $f, f' \in F(X)$. If X is a topological space, then $P_w(X)$ is also endowed with the weak topology generated by $\{\succsim \in P_w(X) \mid f \succ f'\}$ for any continuous $f, f' \in F(X)$.

We will sometimes work with redundant representations of state-dependent preferences in which we distinguish between beliefs and utilities. For a belief $\nu \in \Delta(X)$ and a bounded and measurable utility function $u \colon X \times Z \to \mathbb{R}$ with $u(x,z) \geq u(x,w)$ for all $x \in X$ and $z \in Z$, with strict inequalities for ν-almost every $x \in X$ and some $z \in Z$, we write $\succsim^{\nu,u} \in P_w(X)$ for the induced preference, i.e.,

$$f \succsim^{\nu,u} f' \Leftrightarrow \int_X u(x, f(x)) \, d\nu(x) \geq \int_X u(x, f'(x)) \, d\nu(x)$$

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

4.2 Preference Type Spaces

Fix a finite set $\mathcal{I} = \{1, \ldots, I\}$ of agents with $I \geq 2$ and a compact and metrizable set Θ of states of nature. Each agent i has the worst outcome $w_i \in Z$. We write $P_i(X) \equiv P_{w_i}(X)$.

Definition 4 *A preference type space* $\mathcal{T} = (T_i, \pi_i)_{i \in \mathcal{I}}$ *consists of, for each* $i \in \mathcal{I}$*, a measurable space* T_i *of agent* i*'s types and a measurable mapping* $\pi_i \colon T_i \to P_i(\Theta \times T_{-i})$ *that maps his types to preferences over acts over observable states and his opponents' types, where* $T_{-i} = \prod_{j \in \mathcal{I} \setminus \{i\}} T_j$.

Similarly to Harsanyi type spaces, a product $\tilde{T} = \prod_i \tilde{T}_i$ of measurable sets $\tilde{T}_i \subseteq T_i$ is preference-closed if for every $i \in \mathcal{I}$ and $t_i \in \tilde{T}_i$, $\Theta \times \tilde{T}_{-i}$ is $\pi_i(t_i)$-certain. A type t_i is countable if there exists a countable preference-closed subspace $\tilde{T} = \prod_j \tilde{T}_j$ with $t_i \in \tilde{T}_i$.

For a given Harsanyi type space $\mathcal{T} = \left(\Omega, (T_i, \nu_i, u_i)_{i \in \mathcal{I}}\right)$, we have observed in Section 3.1 two forms of decision theoretic redundancy: first, we can integrate out unobserved states; second, the distinction between beliefs and utilities is not relevant. In particular, a type t_i of agent i is characterized in the Harsanyi type space by a belief $\nu_i(t_i) \in \Delta(\Theta \times \Omega \times T_{-i})$ and a utility function $u_i(t_i) \colon \Theta \times \Omega \times T_{-i} \times Z \to \mathbb{R}$. Together, they induce the preference relation

$$\pi_i^{\nu_i, u_i}(t_i) \equiv \mathrm{mrg}_{\Theta \times T_{-i}} \succsim^{\nu_i(t_i), u_i(t_i)}$$

over $F(\Theta \times T_{-i})$. Thus the preference type space $\mathcal{T} = (T_i, \pi_i^{\nu_i, u_i})_{i \in \mathcal{I}}$ embodies decision theoretically non-redundant information in the Harsanyi type space, and we will abuse notation by writing $\mathcal{T}$ for both when no confusion arises. We will refer to $(T_i, \pi_i^{\nu_i, u_i})_{i \in \mathcal{I}}$ as the preference type space induced by Harsanyi type space $(\Omega, (T_i, \nu_i, u_i)_{i \in \mathcal{I}})$ and refer to types t_i as belonging to both a Harsanyi type space and its induced preference-type space.

4.3 The Universal Preference Type Space

We now construct the universal preference type space à la Mertens and Zamir (1985) and Brandenburger and Dekel (1993). In the light of the isomorphism between preferences $P_i(X)$ and probability measures $\Delta(X \times (Z \setminus \{w_i\}))$ that represent them, this is straightforward and we report standard results with minimal comments.

Let $X_{i,0} = \{*\}$ be initialized with a single element, and let $X_{i,n} = X_{i,n-1} \times P_i(\Theta \times X_{-i,n-1})$ for each $n \geq 1$. Note that $X_{i,n} = \prod_{k=0}^{n-1} P_i(\Theta \times X_{-i,k})$. Let $X_{i,\infty} = \prod_{n=0}^{\infty} P_i(\Theta \times X_{-i,n})$. Each $X_{i,n}$ is compact and metrizable, and thus $X_{i,\infty}$ is compact and metrizable. Let $Y_{i,0} = \prod_{n=0}^{\infty} \Delta(\Theta \times X_{-i,n} \times (Z \setminus \{w_i\}))$ be the set of hierarchies of probability measures for agent i. A hierarchy of probability measures, $\{\mu_{i,n}\}_{n=1}^{\infty} \in Y_{i,0}$, is *coherent* if $\mathrm{mrg}_{\Theta \times X_{-i,n-2} \times (Z \setminus \{w_i\})} \mu_{i,n} = \mu_{i,n-1}$ for every $n \geq 2$. Let $Y_{i,1} \subset Y_{i,0}$ be the set of all coherent hierarchies of probability measures.

For each $\mu_{i,n} \in \Delta(\Theta \times X_{-i,n-1} \times (Z \setminus \{w_i\}))$ with $n \geq 1$, let $\rho_{i,n}(\mu_{i,n}) \in P_i(\Theta \times X_{-i,n-1})$ denote the preference represented by $\mu_{i,n}$. Let $\rho_i \colon Y_{i,0} \to X_{i,\infty}$ be the collection of such mappings $\rho_{i,n}$. Similarly, for each $\mu_{i,\infty} \in \Delta(\Theta \times X_{-i,\infty} \times (Z \setminus \{w_i\}))$, let $\rho_{i,\infty}(\mu_{i,\infty}) \in P_i(\Theta \times X_{-i,\infty})$ denote the preference represented by $\mu_{i,\infty}$.

By the Kolmogorov extension theorem, there is a homeomorphism $\psi_i \colon Y_{i,1} \to \Delta(\Theta \times X_{-i,\infty} \times (Z \setminus \{w_i\}))$. Let $T_{i,1} = \rho_i(Y_{i,1}) \subset X_{i,\infty}$. Note that every $\{\succsim_{i,n}\}_{n=1}^{\infty} \in T_{i,1}$ satisfies coherency, i.e., $\mathrm{mrg}_{\Theta \times X_{-i,n-2}} \succsim_{i,n} = \succsim_{i,n-1}$ for every $n \geq 2$. We convert ψ_i to a mapping between preference spaces and obtain a homeomorphism $\psi_{i,P} = \rho_{i,\infty} \circ \psi_i \circ \rho_i^{-1} \colon T_{i,1} \to P_i(\Theta \times X_{-i,\infty})$.

For $n \geq 2$, let

$$T_{i,n} = \{t_i \in T_{i,1} \mid \Theta \times T_{-i,n-1} \text{ is } \psi_{i,P}(t_i)\text{-certain}\}$$

and $T_i^* = \bigcap_{n=1}^{\infty} T_{i,n}$. Note that $T_{i,n}$ is compact for every $n \geq 1$, and hence T_i^* is also compact. Thus we obtain a homeomorphism $\pi_i^* = \psi_{i,P}|_{T_i^*} \colon T_i^* \to P_i(\Theta \times T_{-i}^*)$. We call $\mathcal{T}^* = (T_i^*, \pi_i^*)_{i \in \mathcal{I}}$ the *universal preference type space*.

Definition 5 *For two preference type spaces $\mathcal{T} = (T_i, \pi_i)_{i \in \mathcal{I}}$ and $\mathcal{T} = (T_i', \pi_i')_{i \in \mathcal{I}}$, a profile $(\varphi_i)_{i \in \mathcal{I}}$ of measurable mappings $\varphi_i \colon T_i \to T_i'$ preserves preferences if*

$$\pi_i' \circ \varphi_i = (\mathrm{id}_\Theta \times \varphi_{-i})^P \circ \pi_i$$

17

for every $i \in \mathcal{I}$.

Fix a preference type space $\mathcal{T} = (T_i, \pi_i)_{i=1,2}$. For each type $t_i \in T_i$ of agent i, let $\hat{\pi}_{i,1}(t_i) = \mathrm{mrg}_\Theta \pi_i(t_i)$ and $\hat{\pi}_{i,n}(t_i) = (\mathrm{id}_\Theta \times (\hat{\pi}_{-i,1}, \ldots, \hat{\pi}_{-i,n-1}))^P(\pi_i(t_i))$ for each $n \geq 2$. Each $\hat{\pi}_{i,n}(t_i)$ denotes the *n-th order preference of t_i*, and $\hat{\pi}_i(t_i) = \{\hat{\pi}_{i,n}(t_i)\}_{n=1}^{\infty}$ the *hierarchy of preferences of t_i*. For any Harsanyi type space, $\mathcal{T} = \left(\Omega, (T_i, \nu_i, u_i)_{i \in \mathcal{I}}\right)$ and $t_i \in T_i$, we also write $\hat{\pi}_i(t_i)$ the *hierarchy of preferences of t_i*, constructed for the induced preference type space $\mathcal{T} = (T_i, \pi_i^{\nu_i, u_i})_{i \in \mathcal{I}}$.

Proposition 2 *For each preference type space $\mathcal{T} = (T_i, \pi_i)_{i \in \mathcal{I}}$, $(\hat{\pi}_i)_{i \in \mathcal{I}}$ is a preference-preserving mapping from $\mathcal{T}$ to the universal type space $\mathcal{T}^*$.*

We write $\hat{\pi}_i(t_i, \mathcal{T})$ for the hierarchy of preferences of t_i when we emphasize the preference type space $\mathcal{T}$ to which t_i belongs.

Definition 6 *Two types t_i in $\mathcal{T}$ and t_i' in $\mathcal{T}'$ have equivalent preference hierarchies if they map to the same type in T_i^*, i.e., $\hat{\pi}_i(t_i, \mathcal{T}) = \hat{\pi}_i(t_i', \mathcal{T}')$.*

5 Equilibrium Strategic Distinguishability

To give a characterization of equilibrium strategic distinguishability, we must require types to be countable in order to ensure existence. Now we have:

Theorem 1 *Two countable types are strategically indistinguishable if and only if they have equivalent preference hierarchies.*

Countability is required only to show the existence of a local equilibrium, and any other set of conditions ensuring existence of a local equilibrium would be sufficient. Proposition 3 below establishes that if two types have equivalent preference hierarchies, then they are strategically indistinguishable. The argument is as follows: suppose agent i expects other agents to follow strategies that are measurable with respect to their higher-order preferences. Then it is a best response to choose a strategy that is measurable with respect to his own higher-order preferences. To show the converse, we will construct a mechanism in which any pair of types that do not have equivalent preference hierarchies have disjoint equilibrium actions. We postpone this proof to Section 6.2.

Lemma 1 *For every pair of type spaces $\mathcal{T}$ and $\mathcal{T}'$, if $\varphi = (\varphi_i)_{i \in \mathcal{I}}$ is a preference-preserving mapping from $\mathcal{T}$ to $\mathcal{T}'$, then $E_i(t_i, \mathcal{T}, \mathcal{M}) \supseteq E_i(\varphi_i(t_i), \mathcal{T}', \mathcal{M})$ for every $i \in \mathcal{I}$, $t_i \in T_i$ and mechanism $\mathcal{M}$.*

Proof. Pick any local equilibrium $\sigma' = (\sigma'_i)$ of $(\mathcal{T}, \mathcal{M})$ associated with preference-closed sub-space $\tilde{T}' = \prod_i T'_i$ of $\mathcal{T}'$. Let $\tilde{T}_i = \varphi_i^{-1}\left(\tilde{T}'_i\right)$ and $\sigma_i = \sigma'_i \circ \varphi_i$. Since φ preserves preferences, $\tilde{T} = \prod_i \tilde{T}_i$ is a preference-closed subspace of $\mathcal{T}$ and $\sigma = (\sigma_i)$ is a local equilibrium of $(\mathcal{T}, \mathcal{M})$ associated with $\tilde{T}$. ∎

Proposition 3 *For two countable types t_i in $\mathcal{T}$ and t'_i in $\mathcal{T}'$ with $\hat{\pi}_i(t_i, \mathcal{T}) = \hat{\pi}_i(t'_i, \mathcal{T}')$, we have $E_i(t_i, \mathcal{T}, \mathcal{M}) \cap E_i(t'_i, \mathcal{T}', \mathcal{M}) \neq \emptyset$ for any mechanism $\mathcal{M}$.*

Proof. By Proposition 2, $\hat{\pi}(\cdot, \mathcal{T})$ and $\hat{\pi}(\cdot, \mathcal{T}')$ are preference-preserving mappings from $\mathcal{T}$ and $\mathcal{T}'$ to the universal space $\mathcal{T}^*$, respectively. By Lemma 1, we have $E_i(t_i, \mathcal{T}, \mathcal{M}) \cap E_i(t'_i, \mathcal{T}', \mathcal{M}) \supseteq E_i(t_i^*, \mathcal{T}^*, \mathcal{M})$, where $t_i^* = \hat{\pi}_i(t_i, \mathcal{T}) = \hat{\pi}_i(t'_i, \mathcal{T}')$. Since t_i is countable in $\mathcal{T}$, t_i^* is also countable in $\mathcal{T}^*$, thus $E_i(t_i^*, \mathcal{T}^*, \mathcal{M}) \neq \emptyset$. ∎

6 Rationalizability

We introduce a natural definition of interim preference correlated rationalizability (IPCR) for the worst outcome preference environments studied in this paper. We then show how our characterization of strategic indistinguishability for equilibrium reported in Theorem 1 continues to hold for this definition of rationalizability. As a corollary, the equivalent preference hierarchies characterize strategic indistinguishability for any solution concept which coarsens equilibrium and refines IPCR. We then report a proof of this result, which will imply the part of Theorem 1 which we did not yet prove.

6.1 Interim Preference Correlated Rationalizability

Fix a preference type space $\mathcal{T} = (T_i, \pi_i)_{i \in \mathcal{I}}$. Write $\Gamma_i \colon T_i \rightrightarrows A_i$ for a correspondence specifying for each type t_i of agent i, a set of actions $\Gamma_i(t_i)$ that are available to type t_i. Fix a profile Γ_{-i} of correspondences of all agents except i. Suppose that agent i were convinced that each agent j of type t_j will choose an action in $\Gamma_j(t_j)$. We will say that action a_i is a best response for t_i against Γ_{-i} if there exists a preference for type t_i in $P_i(\Theta \times T_{-i} \times A_{-i})$ under which (1) there is certainty that action-type profiles of agents other than i are consistent with Γ_{-i}; (2) the marginal preference over $F(\Theta \times T_{-i})$ is consistent with type t_i's original preferences; and (3) a_i is a best response. A correspondence profile $\Gamma = (\Gamma_i)_{i \in \mathcal{I}}$ is a best response correspondence if every action allowed for any type of any agent is a best response to the behavior of other agents. An action is interim preference correlated rationalizable for a given type if it is a possible action for that type in a best response correspondence. More formally:

Definition 7 *Fix a type space $\mathcal{T}$ and a mechanism $\mathcal{M}$. An action $a_i \in A_i$ is a best reply for type $t_i \in T_i$ against Γ_{-i} if there exists $\succsim_i \in P_i(\Theta \times T_{-i} \times A_{-i})$ such that $\Theta \times \mathrm{graph}(\Gamma_{-i})$ is $\succsim_i$-certain, $\mathrm{mrg}_{\Theta \times T_{-i}} \succsim_i = \pi_i(t_i)$ and*

$$\forall a_i' \in A_i, \quad g(\cdot, a_i, \cdot) \ (\mathrm{mrg}_{\Theta \times A_{-i}} \succsim_i) \ g(\cdot, a_i', \cdot).$$

$\Gamma = (\Gamma_i)_{i \in \mathcal{I}}$ is a best reply correspondence if, for every $i \in \mathcal{I}$, $t_i \in T_i$, and $a_i \in \Gamma_i(t_i)$, a_i is a best reply for type t_i against Γ_{-i}. An action a_i is interim preference correlated rationalizable (IPCR) for type t_i if there exists a best reply correspondence Γ with $\Gamma_i(t_i) \ni a_i$.

We write $R_i(t_i, \mathcal{T}, \mathcal{M})$ for the set of IPCR actions for type t_i in type space $\mathcal{T}$ and mechanism $\mathcal{M}$. As usual, we can define $R_i(t_i, \mathcal{T}, \mathcal{M})$ recursively: let $R_{i,0}(t_i, \mathcal{T}, \mathcal{M}) = A_i$ for every $i \in \mathcal{I}$ and $t_i \in T_i$, and, for every $n \geq 1$, let $R_{i,n}(t_i, \mathcal{T}, \mathcal{M})$ be the set of all best replies for type t_i against $R_{-i,n-1}(\cdot, \mathcal{T}, \mathcal{M})$. One can show that $R_i(t_i, \mathcal{T}, \mathcal{M}) = \bigcap_{n \geq 0} R_{i,n}(t_i, \mathcal{T}, \mathcal{M})$, which is nonempty.

IPCR is a very permissive notion of rationalizability. In particular, it allows agents to believe that others' actions convey information about their own preferences over outcomes (consistent with the maintained worst outcome assumption). In the example of Section 3.2, both actions were IPCR even though action D was dominated if one assumed that the opponent's action did not convey payoff relevant information. Morris and Takahashi (2010) show a formal sense in which this definition of rationalizability captures the implications of common certainty of rationality, under the assumption of expected utility preferences that respect worst outcomes.

Definition 8 *Two types of agent i, t_i in $\mathcal{T}$ and t_i' in $\mathcal{T}'$, are IPCR strategically indistinguishable if, for every mechanism $\mathcal{M}$, there exists some action that can be chosen by both types, so that $R_i(t_i, \mathcal{T}, \mathcal{M}) \cap R_i(t_i', \mathcal{T}', \mathcal{M}) \neq \emptyset$ for every $\mathcal{M}$. Conversely, t_i and t_i' are IPCR strategically distinguishable if there exists a mechanism in which no action can be chosen by both types, so that $R_i(t_i, \mathcal{T}, \mathcal{M}^*) \cap R_i(t_i', \mathcal{T}', \mathcal{M}^*) = \emptyset$ for some $\mathcal{M}^*$.*

Theorem 2 *Two types are IPCR strategically indistinguishable if and only if they have equivalent preference hierarchies.*

Under the countability assumption, the direction that if two types are HOP equivalent, then they are IPCR strategically indistinguishable follows from Proposition 3, which proved the corresponding step in Theorem 1, as equilibrium actions are a subset of IPCR actions. However, IPCR actions always exist even for uncountable types. In this case, an analogous argument goes through, but we must appeal to Theorem 3, which shows that two types with equivalent preference hierarchies are IPCR strategically equivalent.

6.2 Proof of Theorems 1 and 2

Let d_i^* be a metric compatible with the product topology on the universal space $T_i^* \subset \prod_{n=0}^{\infty} P_i(\Theta \times X_{-i,n})$. The remaining direction of Theorems 1 and 2 follows from the next proposition.

Proposition 4 *For every $\varepsilon > 0$, there exists a mechanism $\mathcal{M}^*$ such that*

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

for every pair of type spaces $\mathcal{T}$ and $\mathcal{T}'$, $i \in I$, $t_i \in T_i$, and $t_i' \in T_i'$.

Note that Proposition 4 is stronger than necessary to prove Theorems 1 and 2. In particular, the construction of $\mathcal{M}^*$ depends on ε, but is independent of any details of type spaces $\mathcal{T}$ and $\mathcal{T}'$ or any pair of two types t_i and t_i' that we want to distinguish.

Abreu and Matsushima (1992) proved such a result for finite type spaces. In the universal belief type space (the space of Mertens-Zamir hierarchies), Dekel, Fudenberg, and Morris (2006, Lemma 4) construct a discretized direct mechanism in which only actions close to truth telling are interim correlated rationalizable. As we discuss below in Section 7.3, their result corresponds to Proposition 4 under the restriction of common certainty of payoffs. Our proof uses a similar mechanism to both papers, with agents essentially reporting their first level (belief or preference) type, their second level type, and so on. Agents can be given individual incentives to report their first level types truthfully and then inductively, if all agents report their kth level types truthfully, each agent can be given an incentive to report his $(k+1)$th level type truthfully by making outcomes contingent on kth level report of other agents. Two complications may potentially destroy the agents' incentives for truth-telling: (i) Outcomes are not necessarily private goods, and in particular the social planner cannot necessarily give a reward to one agent without affecting the other agents' incentives. Especially, an agent's incentives to report her lower-order preferences are affected by how the social planner uses her reports to solicit other agents' higher-order preferences. (ii) As an agent sends less accurate reports about her lower-order preferences, other agents become less willing to report their higher-order preferences accurately. (i) originates the issue, whereas (ii) "multiplies" it.[2] The finiteness assumption allows Abreu and Matsushima (1992) to deal with both issues by making higher level reports have uniformly lower impact on agents' preferences than lower level reports. Dekel, Fudenberg, and Morris (2006) implicitly assume private goods, removing problem (i). We must carefully exploit our structural assumptions, such as compactness and metrizability of Θ, continuity and monotone continuity of preferences, and existence of the worst outcome, to deal with these issues from the original truth-telling mechanism. The next two subsections are devoted to the proof of Proposition 4.

---

[2]Inaccurate reports may occur in Dekel, Fudenberg, and Morris (2006), but they come purely from discretization.

6.2.1 Single-Agent Revelation Mechanisms

As a preliminary step, here we analyze a single-agent mechanism that reveals her preferences. In this subsection, fix a compact metric space X of states with metric d. Let d_P be a metric compatible with the topology on $P_w(X)$. For each $\succsim \in P_w(X)$, we define the indicator function of $\succsim$, $\chi_{\succsim}$, that maps pairs of acts $f, f' \in F(X)$ to 0, 1/2, or 1 as follows:

$$\chi_{\succsim}(f, f') = \begin{cases} 1 & \text{if } f \succ f', \\ 1/2 & \text{if } f \sim f', \\ 0 & \text{if } f \prec f' \end{cases}$$

for any $f, f' \in F(X)$. 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 $\mathcal{M}^0 = (P_w(X), g^0)$ for a single agent with action set $P_w(X)$ and outcome function

$$g^0(\cdot, a) = \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} 2^{-k-l+1} \chi_a(f_k, f_l) f_k \tag{1}$$

for each $a \in P_w(X)$. Under the mechanism $\mathcal{M}^0$, the agent reports her preference. Then the social planner randomly draws a pair of acts from F and assigns the agent with her preferred act according to her reported preference.[3]

In Lemma 2 below, we show that truth telling is a dominant strategy in $\mathcal{M}^0$ for every type. Indeed, by invoking the compactness of X, we show a "robust" version of strategy proofness: in every mechanism close to $\mathcal{M}^0$, the agent strictly prefers reporting almost true preferences to reporting others according to almost true preferences.

Recall that, for each report $a \in P_w(X)$, $g^0(\cdot, a)$ is an act over X, which determines an outcome z with probability $g^0(x, a)(z)$ when the nature chooses $x \in X$. We consider two sources of perturbations to this act. First, the outcome may not be chosen according to $g^0(x, a)$ with small probability. Formally, for each $\delta > 0$ and measurable space C, we consider perturbed outcome function $g \colon X \times P_w(X) \times C \to \Delta(Z)$ such that $|g(\cdot, \cdot, c) - g^0| = \sup_{x \in X, a \in P_w(X)} |g(x, a, c) - g^0(x, a)| \leq \delta$ for every $c \in C$. Second, nature may choose x' in a neighborhood of x when instead nature is supposed to choose x. Formally, for each $\delta > 0$, let D^δ be the δ-neighborhood of the diagonal of

---

[3]Strictly speaking, $\mathcal{M}^0$ is not a mechanism according to our definition, for its action set is infinite. The mechanism we will construct in the next subsection to prove Proposition 4, however, has finite actions.

$X \times X$, $\{(x, x') \in X \times X \mid d(x, x') \leq \delta\}$. For each $\delta > 0$, $\succsim \in P_w(X)$, and measurable space C, let

$$P_w^{\delta, C}(\succsim) = \left\{ \mathrm{mrg}_{2,3} \succsim' \in P_w(X \times C) \mid \begin{array}{l} \exists \succsim' \in P_w(X \times X \times C) \text{ s.t.} \\ \text{(1) } \mathrm{mrg}_1 \succsim' = \succsim, \\ \text{(2) } D^\delta \times C \text{ is } \succsim'\text{-certain,} \end{array} \right\}, \quad (2)$$

where $\mathrm{mrg}_\Lambda \succsim'$ with $\Lambda \subset \{1, 2, 3\}$ denotes the marginal of $\succsim'$ with respect to the coordinates in Λ. In words, $P_w^{\delta, C}(\succsim)$ is the set of preferences over noisy acts induced by the original preference $\succsim$.

Lemma 2 *For every $\varepsilon > 0$, there exists $\delta > 0$ such that the following holds: for every preference $\succsim \in P_w(X)$, every pair of reports $a, b \in P_w(X)$ that satisfy $d_P(\succsim, a) \leq \delta$ and $d_P(\succsim, b) > \varepsilon$, every measurable space C, and every perturbed outcome function $g \colon X \times P_w(X) \times C \to \Delta(Z)$ that satisfies $|g(\cdot, \cdot, c) - g^0| \leq \delta$ for every $c \in C$, the agent strictly prefers $g(\cdot, a, \cdot)$ to $g(\cdot, b, \cdot)$ according to every preference in $P_w^{\delta, C}(\succsim)$.*

Proof. See Appendix. ∎

6.2.2 Proof of Proposition 4

Let d_Θ be a metric compatible with the topology on Θ. For each $i \in \mathcal{I}$ and $n \geq 1$, let $d_{P,i,n}$ be a metric compatible with the topology on the set of agent i's n-th order preferences, $P_i(\Theta \times X_{-i,n-1})$, and let $d_{i,n}$ be

$$d_{i,n}((\theta, t_{-i,1}, \ldots, t_{-i,n}), (\theta', t'_{-i,1}, \ldots, t'_{-i,n})) = \max \left\{ d_\Theta(\theta, \theta'), \max_{1 \leq k \leq n, j \neq i} d_{P,j,k}(t_{j,k}, t'_{j,k}) \right\},$$

which is a metric compatible with the product topology on $\Theta \times X_{-i,n} = \Theta \times \prod_{k=0}^{n-1} \prod_{j \neq i} P_j(\Theta \times X_{-j,k})$.

Fix any $\varepsilon > 0$. Recall that d_i^* is a metric compatible with the product topology on $T_i^* \subset \prod_{n=0}^\infty P_i(\Theta \times X_{-i,n})$. By the definition of the product topology, there exist $\bar{\varepsilon} > 0$ and $N \in \mathbb{N}$ such that, for every $t_i = \{t_{i,n}\}_{n=1}^\infty$, $t_i' = \{t'_{i,n}\}_{n=1}^\infty \in T_i^*$, if $d_i^*(t_i, t_i') > \varepsilon$, then there exists some $n \leq N$ such that $d_{P,i,n}(t_{i,n}, t'_{i,n}) > \bar{\varepsilon}$. Pick such $\bar{\varepsilon}$ and N.

For each $i \in \mathcal{I}$ and $n \leq N$, substitute $X = \Theta \times X_{-i,n-1}$, $d = d_{i,n-1}$, and $d_P = d_{P,i,n}$ in Section 6.2.1. Pick a countable dense subset of $F_c(\Theta \times X_{-i,n-1})$, and define $g_{i,n}^0 \colon \Theta \times X_{-i,n-1} \times P_i(\Theta \times X_{-i,n-1}) \to \Delta(Z)$ as in (1). For $\delta > 0$, define $D_{i,n}^\delta$ as the δ-neighborhood of the diagonal of $\Theta \times X_{-i,n-1} \times \Theta \times X_{-i,n-1}$. For $\delta > 0$, $\succsim_{i,n} \in P_i(\Theta \times X_{-i,n-1})$, and measurable space C, define $P_{i,n}^{\delta, C}(\succsim_{i,n})$ as in (2). By Lemma 2, there exist $0 < \varepsilon_0 \leq \varepsilon_1 \leq \cdots \leq \varepsilon_{N-1} \leq \varepsilon_N \leq \bar{\varepsilon}/2$ such that, for every $i \in \mathcal{I}$ and $n \leq N$, for every preference $\succsim_{i,n} \in P_i(\Theta \times X_{-i,n-1})$, every pair of reports $a_{i,n}, b_{i,n} \in P_i(\Theta \times X_{-i,n-1})$ that satisfy $d_{P,i,n}(\succsim_{i,n}, a_{i,n}) \leq \varepsilon_{n-1}$ and $d_{P,i,n}(\succsim_{i,n}, b_{i,n}) > \varepsilon_n$, every measurable

23

space C, and every perturbed outcome function $g_{i,n}\colon \Theta \times X_{-i,n-1} \times P_i(\Theta \times X_{-i,n-1}) \times C \to \Delta(Z)$ that satisfies $|g_{i,n}(\cdot,\cdot,c) - g_{i,n}^0| \leq \varepsilon_{n-1}$ for every $c \in C$, player i strictly prefers $g_{i,n}(\cdot, a_{i,n}, \cdot)$ to $g_{i,n}(\cdot, b_{i,n}, \cdot)$ according to every preference in $P_{i,n}^{\varepsilon_{n-1},C}(\succsim_{i,n})$.

We define a mechanism $\mathcal{M}^* = ((A_i^*)_{i \in \mathcal{I}}, g^*)$ as follows. For each $i \in \mathcal{I}$ and $n \leq N$, let $A_{i,n}^*$ be any ε_{n-1}-dense finite subset of $P_i(\Theta \times X_{-i,n-1})$ with respect to $d_{P,i,n}$, and $A_i^* = \prod_{n=1}^N A_{i,n}^*$. Define $g^*\colon \Theta \times A^* \to \Delta(Z)$ by

$$g^*(\theta, a) = \frac{1-\delta}{I(1-\delta^N)} \sum_{i=1}^I \sum_{n=1}^N \delta^{n-1} g_{i,n}^0(\theta, a_{-i,1}, \ldots, a_{-i,n-1}, a_{i,n})$$

for each $\theta \in \Theta$ and $a = (a_{i,n}) \in A^*$, where $\delta > 0$ is small enough to satisfy $(1-\delta)/\delta \geq (I-1)(1-\varepsilon_0)/\varepsilon_0$.

Claim 1 *For every type space $\mathcal{T} = (T_i, \pi_i)_{i \in \mathcal{I}}$ and $n \leq N$, we have*

$$a_i \in R_{i,n}(t_i, \mathcal{T}, \mathcal{M}^*) \Rightarrow d_{P,i,n}(\hat{\pi}_{i,n}(t_i, \mathcal{T}), a_{i,n}) \leq \varepsilon_n$$

for every $i \in \mathcal{I}$ and $t_i \in T_i$.

Proof. The proof is by induction on n. Suppose that, for every $k \leq n-1$, $a_i \in R_{i,n-1}(t_i, \mathcal{T}, \mathcal{M}^*)$ implies $d_{P,i,k}(\hat{\pi}_{i,k}(t_i, \mathcal{T}), a_{i,k}) \leq \varepsilon_k \leq \varepsilon_{n-1}$ for every $i \in \mathcal{I}$ and $t_i \in T_i$. Suppose that there exists $a_i^* \in R_{i,n}(t_i, \mathcal{T}, \mathcal{M}^*)$ such that $d_{P,i,n}(\hat{\pi}_{i,n}(t_i, \mathcal{T}), a_{i,n}^*) > \varepsilon_n$. Then there exists $\succsim_i \in P_i(\Theta \times T_{-i} \times A_{-i}^*)$ such that $\Theta \times \mathrm{graph}(R_{-i,n-1}(\cdot, \mathcal{T}, \mathcal{M}^*))$ is $\succsim_i$-certain, $\mathrm{mrg}_{\Theta \times T_{-i}} \succsim_i = \pi_i(t_i)$, and player i weakly prefers $g^*(\cdot, a_i^*, \cdot)$ to $g^*(\cdot, a_i, \cdot)$ for every $a_i \in A_i^*$ according to $\mathrm{mrg}_{\Theta \times A_{-i}^*} \succsim_i$.

Let $C = \prod_{k=n}^N A_{-i,k}^*$ and $\varphi_{-i}\colon \Theta \times T_{-i} \times A_{-i}^* \to \Theta \times X_{-i,n-1} \times \Theta \times X_{-i,n-1} \times C$ such that $\varphi_{-i}(\theta, t_{-i}, a_{-i}) = (\theta, \hat{\pi}_{-i,1}(t_{-i}, \mathcal{T}), \ldots, \hat{\pi}_{-i,n-1}(t_{-i}, \mathcal{T}), \theta, a_{-i,1}, \ldots, a_{-i,n-1}, a_{-i,n}, \ldots, a_{-i,N})$. Collect all the terms in g^* that depend on $a_{i,n}$ and define $g_{i,n}^*\colon \Theta \times X_{-i,n-1} \times A_{i,n}^* \times C \to \Delta(Z)$ by

$$g_{i,n}^*(\theta, a_{-i,1}, \ldots, a_{-i,n-1}, a_{i,n}, a_{-i,n}, \ldots, a_{-i,N})$$
$$= K\left(g_{i,n}^0(\theta, a_{-i,1}, \ldots, a_{-i,n-1}, a_{i,n}) + \sum_{j \in \mathcal{I} \setminus \{i\}} \sum_{k=n+1}^N \delta^{k-n} g_{j,k}^0(\theta, a_{-j,1}, \ldots, a_{-j,k-1}, a_{j,k}) \right),$$

where $a_{i,k} = a_{i,k}^*$ for $k \neq n$ when they appear in the second term, and K is a positive normalization constant. Since we chose sufficiently small δ, we have $|g_{i,n}^*(\cdot, \cdot, c) - g_{i,n}^0| \leq \varepsilon_0 \leq \varepsilon_{n-1}$ for every $c \in C$. Let $\succsim_i' = (\varphi_{-i})^P(\succsim_i)$. By the induction hypothesis, $\varphi_{-i}(\Theta \times \mathrm{graph}(R_{-i,n-1}(\cdot, \mathcal{T}, \mathcal{M}^*))) \subseteq D_{i,n}^{\varepsilon_{n-1}} \times C$ is $\succsim_i'$-certain. Thus, we have $\mathrm{mrg}_{\Theta \times A_{-i}^*} \succsim_i \in P_{i,n}^{\varepsilon_{n-1},C}(\hat{\pi}_{i,n}(t_i, \mathcal{T}))$. Since $A_{i,n}^*$ is ε_{n-1}-dense in $P_i(\Theta \times X_{-i,n-1})$, there exists $a_{i,n}' \in A_{i,n}^*$ such that $d_{P,i,n}(\hat{\pi}_{i,n}(t_i, \mathcal{T}), a_{i,n}') \leq \varepsilon_{n-1}$. By

24

Lemma 2, $\text{mrg}_{\Theta \times A^*_{-i}} \succsim_i$ strictly prefers $g^*_{i,n}(\cdot, a'_{i,n}, \cdot)$ to $g^*_{i,n}(\cdot, a^*_{i,n}, \cdot)$, thus $\text{mrg}_{\Theta \times A^*_{-i}} \succsim_i$ strictly prefers $g^*(\cdot, a'_{i,n}, a^*_{i,-n}, \cdot)$ to $g^*(\cdot, a^*_i, \cdot)$. This is a contradiction. ∎

We can now complete the proof of Proposition 4.

Proof of Proposition 4. Pick any pair of type spaces $\mathcal{T}$ and $\mathcal{T}'$, $i \in \mathcal{I}$, $t_i \in T_i$, and $t'_i \in T'_i$. Suppose that there exists $a_i = (a_{i,1}, \ldots, a_{i,N}) \in R_i(t_i, \mathcal{T}, \mathcal{M}^*) \cap R_i(t'_i, \mathcal{T}', \mathcal{M}^*)$. 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

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

by Claim 1. Thus $d^*_i(\hat{\pi}_i(t_i, \mathcal{T}), \hat{\pi}_i(t'_i, \mathcal{T}')) \leq \varepsilon$. ∎

7   Rationalizability and Strategic Equivalence

Our notion of strategic distinguishability is very demanding: in some game, two types have no equilibrium (or rationalizable) actions in common. The notion of strategic indistinguishability is correspondingly undemanding: it is enough that the two types have any equilibrium (or rationalizable) action in common in some game. In this section, we will study the alternative notion of strategic equivalence. Two types are strategically equivalent if they have the same set of equilibrium (or rationalizable) actions. For any (nonempty-valued) solution concept, strategic equivalence is a stronger requirement than strategic indistinguishability and thus implies a finer partition of types. The corresponding notion of strategic non-equivalence will then be easier to satisfy than strategic distinguishability.

While the characterization of strategic distinguishability is the same for most solution concepts, i.e., for equilibrium, a very permissive definition of rationalizability and everything in between, we will see that strategic equivalence characterizations are sensitive to the solution concept. To understand strategic equivalence and its sensitivity, it is useful to introduce a family of rationalizability notions refining interim preference correlated rationalizability, which impose restrictions on the preferences supporting a best response. Our definition of IPCR allows agents' ex post preferences over lotteries, conditional on others' actions and types, to be anything consistent with the worst outcome assumption. Suppose that we impose a further restriction on agents' possible ex post preferences. A given restriction gives rise to a definition of rationalizability, where preferences supporting a best response must have ex post preferences consistent with the restriction. We show that if we restrict attention to types that belong to type spaces where a given preference restriction holds, then two types are strategically equivalent under the version of rationalizability satisfying that restriction if and only if they have equivalent preference hierarchies.

This result has two important special cases. First, if no restrictions other than the worst outcome assumption are imposed on rationalizability, i.e., if we stick to our earlier definition of IPCR, then this result implies that two types are IPCR strategically equivalent if and only if they have equivalent preference hierarchies. Second, if we impose the restriction that ex post preferences are fixed, i.e., there is common certainty of payoffs, then this result reduces to (a generalization of) the result of Dekel, Fudenberg and Morris (2006, 2007) that, with common certainty of payoffs as a maintained assumption, two types have the same interim correlated rationalizable actions if and only if they have the same Mertens-Zamir higher-order belief hierarchy.

7.1 Ex Post Preference Restrictions

For each $\succsim \in P_w(X)$ and measurable $E \subseteq X$, we write $\succsim_E$ for the *conditional preference* over lotteries defined by

$$y \succsim_E y' \Leftrightarrow y_E y'' \succsim y'_E y''$$

for any $y, y' \in \Delta(Z)$ and some $y'' \in \Delta(Z)$. By independence of $\succsim$, the choice of y'' does not affect the definition of $\succsim_E$.

An *ex post restriction* on agents' preferences will specify a set of possible conditional preferences for each agent. Thus $\mathbf{U} = (U_i)_{i \in I}$, where each U_i is a non-empty set of linearly independent vectors in $\Delta(Z \setminus \{w_i\}) \subset \mathbb{R}^{Z \setminus \{w_i\}}$.[4] The interpretation is that we will impose the requirement that agent i's preferences are representable by convex combinations of U_i, even if they are conditioned on observable states and other agents' types and actions.

We will say that agent i's preference relation $\succsim_i \in P_i(X)$ is U_i-*consistent* if, for any non-$\succsim_i$-null event $E \subseteq X$, the conditional preference $\succsim_{i,E}$ is represented by a convex combination of U_i. A type space $\mathcal{T} = (T_i, \pi_i)_{i \in \mathcal{I}}$ is $\mathbf{U}$-*consistent* if, for each $i \in \mathcal{I}$ and $t_i \in T_i$, $\pi_i(t_i)$ is U_i-consistent. A type t_i is $\mathbf{U}$-*consistent* if it belongs to a $\mathbf{U}$-consistent preference-closed subspace.

We can now define a family of rationalizability concepts for a game $(\mathcal{T}, \mathcal{M})$ on ex post preference restrictions.

Definition 9 *Fix a type space $\mathcal{T}$ and a mechanism $\mathcal{M}$. An action $a_i \in A_i$ is a U_i-best reply for type $t_i \in T_i$ against Γ_{-i} if there exists $\succsim_i \in P_i(\Theta \times T_{-i} \times A_{-i})$ such that $\succsim_i$ is U_i-consistent, $\Theta \times \mathrm{graph}(\Gamma_{-i})$ is $\succsim_i$-certain, $\mathrm{mrg}_{\Theta \times T_{-i}} \succsim_i = \pi_i(t_i)$ and*

$$\forall a'_i \in A_i, \quad g(\cdot, a_i, \cdot) \ (\mathrm{mrg}_{\Theta \times A_{-i}} \succsim_i) \ g(\cdot, a'_i, \cdot).$$

---

[4]Linear independence is a condition imposed on utility representations, but, given the isomorphism between $P_i(\{*\})$ and $\Delta(Z \setminus \{w_i\})$, one can provide an equivalent condition on preferences over lotteries. For more details, see Morris and Takahashi (2010).

$\Gamma = (\Gamma_i)_{i \in \mathcal{I}}$ is a **U**-*best reply correspondence if, for every* $i \in \mathcal{I}$, $t_i \in T_i$, *and* $a_i \in \Gamma_i(t_i)$, a_i *is a* U_i-*best reply for type* t_i *against* Γ_{-i}. *An action* a_i *is interim* **U**-*rationalizable for type* t_i *if there exists a* **U**-*best reply correspondence* Γ *with* $\Gamma_i(t_i) \ni a_i$.

Let $R_i^{\mathbf{U}}(t_i, \mathcal{T}, \mathcal{M})$ be the set of **U**-rationalizable actions for type t_i in game $(\mathcal{T}, \mathcal{M})$. Let $R_{i,0}^{\mathbf{U}}(t_i, \mathcal{T}, \mathcal{M}) = A_i$ for every $i \in \mathcal{I}$ and $t_i \in T_i$, and, for every $n \geq 1$, let $R_{i,n}^{\mathbf{U}}(t_i, \mathcal{T}, \mathcal{M})$ be the set of U_i-best replies for type t_i against $R_{-i,n-1}^{\mathbf{U}}(\cdot, \mathcal{T}, \mathcal{M})$. We have $R_i^{\mathbf{U}}(t_i, \mathcal{T}, \mathcal{M}) = \bigcap_{n \geq 0} R_{i,n}^{\mathbf{U}}(t_i, \mathcal{T}, \mathcal{M})$. Note that $R_i^{\mathbf{U}}(t_i, \mathcal{T}, \mathcal{M})$ is non-empty if and only if t_i is **U**-consistent.

Battigalli and Siniscalchi (2003) define a family of definitions of rationalizability, called "Δ-rationalizability", by imposing restrictions on first order beliefs within the solution concept. "Payoffs" are not incorporated in their type spaces and thus they implicitly maintain common certainty of payoffs over outcomes. **U**-rationalizability parallels Δ-rationalizability in imposing restrictions within the solution concept on beliefs/preferences, but these restrictions concern conditional preferences rather than interim beliefs.

We are most interested in two rationalizable notions, which correspond to the minimal and maximal conditional preference restrictions, respectively. For the minimal case, we have $U_i = \{\bar{u}_i\}$, a singleton, for each agent i. The solution concept $R^{\mathbf{U}}$ then corresponds to "interim correlated rationalizability" with the restriction that agent i's preferences over lotteries are always represented by $\bar{u}_i$. We will discuss this case in detail in Section 7.3. For the maximal case, we have $U_i = \{u_{i,z} \mid z \in Z \setminus \{w_i\}\}$, where $u_{i,z}$ is the unit vector with 1 on outcome z, thus the convex hull of U_i is equal to $\Delta(Z \setminus \{w_i\})$. Then conditional preference restrictions become vacuous, and interim **U**-rationalizability corresponds to IPCR.

Definition 10 *Two types of agent* i, t_i *in* $\mathcal{T}$ *and* t_i' *in* $\mathcal{T}'$, *are* $R^{\mathbf{U}}$ *strategically indistinguishable if, for every mechanism* $\mathcal{M}$, *there exists some action that can be chosen by both types, so that* $R_i^{\mathbf{U}}(t_i, \mathcal{T}, \mathcal{M}) \cap R_i^{\mathbf{U}}(t_i', \mathcal{T}', \mathcal{M}) \neq \emptyset$ *for every* $\mathcal{M}$. *Conversely,* t_i *and* t_i' *are* $R^{\mathbf{U}}$ *strategically distinguishable if there exists a mechanism in which no action can be chosen by both types, so that* $R_i^{\mathbf{U}}(t_i, \mathcal{T}, \mathcal{M}^*) \cap R_i^{\mathbf{U}}(t_i', \mathcal{T}', \mathcal{M}^*) = \emptyset$ *for some* $\mathcal{M}^*$.

An immediate corollary of Theorems 1 and 2 is:

Corollary 1 *For any conditional preference restrictions* **U**, *two* **U**-*consistent types are* $R^{\mathbf{U}}$ *strategically indistinguishable if and only if they have equivalent preference hierarchies.*

7.2 Strategic Equivalence

We introduced the notion of strategic equivalence in Section 3.2. A formal definition is as follows:

Definition 11 *Two types of agent i, t_i in $\mathcal{T}$ and t_i' in $\mathcal{T}'$, are $R^{\mathbf{U}}$ strategically equivalent if, for every mechanism $\mathcal{M}$, $R_i^{\mathbf{U}}(t_i, \mathcal{T}, \mathcal{M}) = R_i^{\mathbf{U}}(t_i', \mathcal{T}', \mathcal{M})$ for every $\mathcal{M}$.*

Now we have:

Theorem 3 *For any conditional preference restrictions $\mathbf{U}$, two $\mathbf{U}$-consistent types are $R^{\mathbf{U}}$ strategically equivalent if and only if they have equivalent preference hierarchies.*

We report a proof for finite type spaces. The proof is close to the proof of Proposition 1 of Dekel, Fudenberg and Morris (2007) and the proof for general type spaces mirrors the proof of Lemma 1 of Dekel, Fudenberg and Morris (2007), the extension of Proposition 1 to general type spaces.

Proof. We will establish by induction on $n \geq 1$ that, if $\hat{\pi}_{i,n}(t_i, \mathcal{T}) = \hat{\pi}_{i,n}(t_i', \mathcal{T}')$, then $R_{i,n}^{\mathbf{U}}(t_i, \mathcal{T}, \mathcal{M}) = R_{i,n}^{\mathbf{U}}(t_i', \mathcal{T}', \mathcal{M})$. Suppose that this holds for $n-1$, that $\hat{\pi}_{i,n}(t_i, \mathcal{T}) = \hat{\pi}_{i,n}(t_i', \mathcal{T}')$ and that $a_i \in R_{i,n}^{\mathbf{U}}(t_i, \mathcal{T}, \mathcal{M})$. Let $\mu_i \in \Delta(\Theta \times T_{-i} \times U_i)$ and $\mu_i' \in \Delta(\Theta \times T_{-i}' \times U_i)$ be probability measures that represent $\pi_i(t_i)$ and $\pi_i'(t_i')$, respectively. Since a_i is a U_i-best reply for t_i against $R_{-i,n-1}^{\mathbf{U}}(\cdot, \mathcal{T}, \mathcal{M})$ in $(\mathcal{T}, \mathcal{M})$, there exists $\nu_i \in \Delta(\Theta \times T_{-i} \times A_{-i} \times U_i)$ such that:

$\quad$ (1) $\quad \nu_i(\theta, t_{-i}, a_{-i}, u_i) > 0 \Rightarrow a_{-i} \in R_{-i,n-1}^{\mathbf{U}}(t_{-i}, \mathcal{T}, \mathcal{M})$,

$\quad$ (2) $\quad \displaystyle\sum_{a_{-i} \in A_{-i}} \nu_i(\theta, t_{-i}, a_{-i}, u_i) = \mu_i(\theta, t_{-i}, u_i)$ for all $\theta \in \Theta, t_{-i} \in T_{-i}, u_i \in U_i$,

$\quad$ (3) $\quad a_i \in \displaystyle\arg\max_{a_i' \in A_i} \sum_{\theta, t_{-i}, a_{-i}, u_i, z} g(\theta, (a_i', a_{-i}))(z) \nu_i(\theta, t_{-i}, a_{-i}, u_i) u_i(z)$.

Let
$$D_{-i,n-1} = \left\{\hat{\pi}_{-i,n-1}(t_{-i}, \mathcal{T}) \mid t_{-i} \in T_{-i}\right\}.$$

For $\hat{\pi}_{-i,n-1} \in D_{-i,n-1}$, let
$$\hat{\mu}_i(\theta, \hat{\pi}_{-i,n-1}, u_i) = \sum_{\hat{\pi}_{-i,n-1}(t_{-i}, \mathcal{T}) = \hat{\pi}_{-i,n-1}} \mu_i(\theta, t_{-i}, u_i).$$

Since $\hat{\pi}_{i,n}(t_i, \mathcal{T}) = \hat{\pi}_{i,n}(t_i', \mathcal{T}')$, μ_i and μ_i' represent the same n-th order preference. Since U_i is linearly independent, $\mu_i = \mu_i'$ induce the same probability distribution over $\Theta \times D_{-i,n-1} \times U_i$, i.e.,
$$\hat{\mu}_i(\theta, \hat{\pi}_{-i,n-1}, u_i) = \sum_{\hat{\pi}_{-i,n-1}(t_{-i}', \mathcal{T}') = \hat{\pi}_{-i,n-1}} \mu_i'(\theta, t_{-i}', u_i)$$

for all $\theta \in \Theta$, $\hat{\pi}_{-i,n-1} \in D_{-i,n-1}$ and $u_i \in U_i$.

For each $(\theta, \hat{\pi}_{-i,n-1}, u_i)$ such that $\hat{\mu}_i(\theta, \hat{\pi}_{-i,n-1}, u_i) > 0$, set
$$\sigma_{-i}(a_{-i} | \theta, \hat{\pi}_{-i,n-1}, u_i) = \frac{1}{\hat{\mu}_i(\theta, \hat{\pi}_{-i,n-1}, u_i)} \sum_{\hat{\pi}_{-i,n-1}(t_{-i}, \mathcal{T}) = \hat{\pi}_{-i,n-1}} \nu_i(\theta, t_{-i}, a_{-i}, u_i).$$

Note that, for each (θ, t_{-i}, u_i) such that $\hat{\mu}_i(\theta, \hat{\pi}_{-i,n-1}(t_{-i}, \mathcal{T}), u_i) > 0$, we have $\sigma_{-i}(a_{-i}|\theta, \hat{\pi}_{-i,n-1}(t_{-i}, \mathcal{T}), u_i) > 0$ only if $a_{-i} \in R^{\mathbf{U}}_{-i,n-1}(t_{-i}, \mathcal{T}, \mathcal{M})$.

Let

$$\nu'_i\left(\theta, t'_{-i}, a_{-i}, u_i\right) = \mu'_i(\theta, t'_{-i}, u_i)\sigma_{-i}\left(a_{-i}|\theta, \hat{\pi}_{-i,n-1}(t'_{-i}, \mathcal{T}'), u_i\right).$$

Note that ν'_i is well defined because, whenever $\mu'_i(\theta, t'_{-i}, u_i) > 0$, we have $\hat{\mu}_i(\theta, \hat{\pi}_{-i,n-1}(t'_{-i}, \mathcal{T}'), u_i) > 0$.

Now we show that a_i is a U_i-best reply for t_i against $R_{-i,n-1}(\cdot, \mathcal{T}', \mathcal{M})$ in $(\mathcal{T}', \mathcal{M})$. First, suppose that $\nu'_i\left(\theta, t'_{-i}, a_{-i}, u_i\right) > 0$. Then there exists $t_{-i} \in T_{-i}$ such that $\hat{\pi}_{-i,n-1}(t_{-i}, \mathcal{T}) = \hat{\pi}_{-i,n-1}(t'_{-i}, \mathcal{T}')$. Since we have $\hat{\mu}_i\left(\theta, \hat{\pi}_{-i,n-1}(t_{-i}, \mathcal{T}), u_i\right) = \hat{\mu}_i\left(\theta, \hat{\pi}_{-i,n-1}(t'_{-i}, \mathcal{T}'), u_i\right) > 0$ and $\sigma_{-i}(a_{-i}|\theta, \hat{\pi}_{-i,n-1}(t_{-i}, \mathcal{T}), u_i) = \sigma_{-i}\left(a_{-i}|\theta, \hat{\pi}_{-i,n-1}(t'_{-i}, \mathcal{T}'), u_i\right) > 0$, we have $a_{-i} \in R^{\mathbf{U}}_{-i,n-1}(t_{-i}, \mathcal{T}, \mathcal{M})$, which is equal to $R^{\mathbf{U}}_{-i,n-1}(t'_{-i}, \mathcal{T}', \mathcal{M})$ by the induction hypothesis.

Second, by the construction of ν'_i, the marginal distribution of ν'_i over $\Theta \times T_{-i} \times U_i$ is equal to μ'_i, which represents $\pi'_i(t'_i)$.

Third, since we have

$$\sum_{t'_{-i}} \nu'_i\left(\theta, t'_{-i}, a_{-i}, u_i\right) = \sum_{t'_{-i}} \mu'_i(\theta, t'_{-i}, u_i)\sigma_{-i}\left(a_{-i}|\theta, \hat{\pi}_{-i,n-1}(t'_{-i}, \mathcal{T}'), u_i\right)$$

$$= \sum_{\hat{\pi}_{-i,n-1} \in D_{-i,n-1}} \hat{\mu}_i\left(\theta, \hat{\pi}_{-i,n-1}, u_i\right)\sigma_{-i}\left(a_{-i}|\theta, \hat{\pi}_{-i,n-1}, u_i\right)$$

$$= \sum_{t_{-i}} \mu_i(\theta, t_{-i}, u_i)\sigma_{-i}\left(a_{-i}|\theta, \hat{\pi}_{-i,n-1}(t_{-i}, \mathcal{T}), u_i\right)$$

$$= \sum_{t_{-i}} \nu_i\left(\theta, t_{-i}, a_{-i}, u_i\right),$$

ν_i and ν'_i have the same marginal distribution over $\Theta \times A_{-i} \times U_i$. Thus a_i is a best reply with respect to ν'_i in $(\mathcal{T}', \mathcal{M})$. $\blacksquare$

Since IPCR corresponds to vacuous conditional preference restrictions, an immediate corollary is:

Corollary 2 *Two types are IPCR strategically equivalent if and only if they have equivalent preference hierarchies.*

7.3 Common Certainty of "Payoffs"

DFM (2006, 2007) show a strategic equivalence result for the solution concept of ICR. In particular, they consider "games" $\mathcal{G} = \left((A_i)_{i \in \mathcal{I}}, \hat{g}\right)$, where A_i is a finite action set for agent i, and a measurable function $\hat{g}: \Theta \times A \to [0,1]^I$ describes "payoffs" as a function of observable states Θ and action

29

profiles. "Payoffs" in correspond to von-Neumann-Morgenstern indices in our setting, and since the function $\hat{g}$ is taken to be common certainty among the agents, it is implicitly assumed that there is common certainty of "payoffs" or von-Neumann-Morgenstern indices. DFM show that two types have the same set of interim correlated rationalizable actions in all games $\mathcal{G}$ if and only if they have the same MZ hierarchy of beliefs and higher-order beliefs about Θ. In particular, Lemma 4 of DFM (2006) establishes that types with distinct MZ hierarchies must have distinct ICR actions; Proposition 1 (for finite type spaces) and Lemma 1 (for infinite type spaces) of DFM (2007) establish that types with the same MZ hierarchy have the same set of ICR actions.

Lemma 4 of DFM (2006) is a special case of our Proposition 4. To see why, let $Z = \prod_i Z_i$ with $Z_i = \{0,1\}$, and $U_i = \{\overline{u}_i\}$ with $\overline{u}_i(z_1, \ldots, z_I) = z_i$. In this case, any belief type space $\mathcal{T} = (T_i, \mu_i)_{i \in \mathcal{I}}$ with $\mu_i \colon T_i \to \Delta(\Theta \times T_{-i})$ induces a preference type space $\mathcal{T}' = (T_i, \pi_i)_{i \in \mathcal{I}}$ by $\pi_i(t_i) = \succsim^{\mu_i(t_i), \overline{u}_i}$. Then IPCR in $(\mathcal{T}', \mathcal{M})$ is more permissive than **U**-rationalizability in $(\mathcal{T}', \mathcal{M})$, which reduces to ICR in $(\mathcal{T}, \mathcal{G})$ (as defined in DFM (2006, 2007)) in the game $\mathcal{G} = \left((A_i)_{i \in \mathcal{I}}, \hat{g} \right)$ with

$$\hat{g}_i(\theta, a) = \sum_z g(\theta, a)(z) \overline{u}_i(z) = \sum_{z_{-i}} g(\theta, a)(1, z_{-i}).$$

Thus our Proposition 4 implies Lemma 4 of DFM (2006).[5] Similarly, Proposition 1 and Lemma 1 of DFM (2007) are a special case of our Theorem 3.

Examples in DFM (2007) and Ely and Pęski (2006) show that under less permissive versions of rationalizability—for example, IIR in DFM (2007)—MZ types do not characterize strategic equivalence. Ely and Pęski (2006) provide a characterization of strategic equivalence for IIR in two-player games. Liu (2009) and Sadzik (2010) provide characterizations of "redundant" components required for equilibrium strategic equivalence. Thus the message of this "common certainty of payoffs" literature is that strategic equivalence is sensitive to the solution concept considered. Although the point was not highlighted in this literature, it is easy to see that Mertens-Zamir higher-order beliefs characterize strategic distinguishability in this common certainty of payoffs setting. Our Corollary 1 makes this point without common certainty of payoffs.

Thus there is a clean parallel between results for the two environments of "common certainty of payoffs" literature and the general case studied in this paper. Independent of the solution concept, strategic distinguishability is characterized by MZ higher-order beliefs and higher-order preferences, respectively. Characterizations of strategic equivalence depend on the solution concept.

---

[5]Indeed, one can show from our Proposition 4 that MZ hierarchies characterize strategic distinguishability even if restrictions are imposed on payoffs across agents. That is, one can use only measurable functions $\hat{g} \colon \Theta \times A \to V$ to strategically distinguish distinct MZ types, where V is a convex subset of $[0,1]^I$ such that, for any agent i, there exist $v^i, \tilde{v}^i \in V$ such that $v_i^i \neq \tilde{v}_i^i$.

ICR strategic equivalence is characterized by MZ higher-order beliefs, and IPCR strategic equivalence is characterized by higher-order preferences. More refined solution concepts may require finer descriptions of types to characterize strategic equivalence.

8 Discussions

8.1 Relaxing the Worst Outcome Property

We have assumed so far that, for each agent i, there is common certainty that an outcome w_i is worse than any other outcome for that agent. There are two roles which the worst outcome assumption plays in our analysis. First, combined with the non-triviality assumption, it rules out the possibility of types that are completely indifferent between all outcomes. Second, it ensures the space $P_w(X)$ of all possible preferences is isomorphic to $\Delta(X \setminus (Z \setminus \{w\}))$, which is compact and metrizable if X is compact and metrizable. Both results are indispensable for our results. Clearly, every action is rationalizable for a completely indifferent type and thus such a type cannot be strategically distinguished from any other type. Also, we can show that—even after ruling out complete indifference—if the set of all possible preferences is not compact, then not only do technical difficulties arise in the construction of a universal preference type space, but more importantly, it is no longer the case that two types with distinct preference hierarchies can be strategically distinguished. This point is discussed in Morris and Takahashi (2010) and is related to the negative results in Ledyard (1986).

The worst outcome assumption is a convenient way of ruling out complete indifference and guaranteeing compactness of the space of possible preferences. However, weaker assumptions will work as well. For $\lambda \in (0, 1/2]$, we say that a binary relation $\succsim$ over $F(X)$ is λ-*continuous* if there exist two outcomes $z, z' \in Z$ with $z \succ z'$ and, for every $f, f' \in F(X)$, we have

$$(1 - \lambda) z + \lambda f \succsim (1 - \lambda) z' + \lambda f'.$$

For a general state-dependent preference, preferences over outcomes may depend on states. λ-continuity requires that the strength of such state dependency be bounded in the sense that, even if an agent receives state-dependent acts with probability λ, it does not alter her preference between state-independent outcomes z and z'.

The notion of λ-continuity is a weak requirement. To see this, note that every binary relation $\succsim$ over $F(X)$ that satisfies completeness, transitivity, independence, continuity and monotone continuity is represented by a finite signed measure μ on $X \times Z$:

$$f \succsim f' \Leftrightarrow \int_{X \times Z} f(x)(z) d\mu(x, z) \geq \int_{X \times Z} f'(x)(z) d\mu(x, z).$$

If a preference is not indifferent over lotteries, then it is λ-continuous for a sufficiently small $\lambda > 0$. For example, one can take $\lambda > 0$ such that

$$\frac{\lambda}{1 - \lambda} \leq \frac{\|\mathrm{mrg}_Z \mu\|}{\|\mu\|},$$

where $\mathrm{mrg}_Z \mu$ is the marginal of μ on Z given by $(\mathrm{mrg}_Z \mu)(\{z\}) := \mu(X \times \{z\})$, and, for $\nu = \mu, \mathrm{mrg}_Z \mu$, $\|\nu\| := \sup_{E, E'} (\nu(E) - \nu(E'))$ (E and E' vary over all measurable sets) denotes the total variation of ν. Also, every preference in $P_w(X)$ is λ-continuous with any $0 < \lambda \leq 1/|Z|$.

Then we focus on preference type spaces where there is common certainty that all agents' preferences are λ-continuous for some fixed $\lambda > 0$. Such spaces include preference type spaces with the worst outcome property, studied in the body of this paper, and other settings, such as finite type spaces of Abreu and Matsushima (1992) and "compact and continuous" type spaces (see Proposition 6 in Appendix C).

For such preference type spaces, we can construct a universal preference type space, consisting of coherent hierarchies of preferences, for each $\lambda > 0$. Also, we can show Theorems 1 and 2, i.e., the universal space characterizes strategic distinguishability for equilibrium, interim preference correlated rationalizability that respects λ-continuity, and everything in between.[6] Details are given in Appendix C.

8.2 Payoff Type Environments

As part of an analysis of robust virtual implementation, Bergemann and Morris (2009) analyze a variant of the strategic distinguishability question. Consider a payoff type environment, where there is a finite set Z of outcomes and a finite set of agents, $\mathcal{I} = \{1, ..., I\}$, each with a payoff type φ_i drawn from a finite set Φ_i and a (perhaps interdependent) utility function $\hat{u}_i \colon \Phi \times Z \to \mathbb{R}$. Common certainty of utility functions $(\hat{u}_i)_{i \in \mathcal{I}}$ - and thus agents' ex post preference conditional on the profile of known ϕ - is (implicitly) assumed. Now a type space $\mathcal{T} = (T_i, b_i, \tilde{\varphi}_i)_{i \in \mathcal{I}}$ specifies for each agent i a set of possible types T_i, and mappings $b_i \colon T_i \to \Delta(T_{-i})$ and $\tilde{\varphi}_i \colon T_i \to \Phi_i$ identifying the beliefs and (known) own payoff type of each types. Expressing the strategic distinguishability question of Bergemann and Morris (2009) in the language of this paper, we can identify the set of (say) equilibrium actions $E_i(t_i, \mathcal{T}, \mathcal{M})$ of type t_i from type space $\mathcal{T}$ playing mechanism $\mathcal{M}$. Now say that payoff types φ_i and φ_i' are strategically distinguishable if there exists a mechanism where—whatever their beliefs and higher-order beliefs about other agents' payoff types—they have no action in common; formally, if there exists a mechanism $\mathcal{M}^*$ such that for all t_i in type space $\mathcal{T}$

---

[6]We do not expect to have a strategic equivalence result such as Theorem 3

with $\tilde{\varphi}_i(t_i) = \varphi_i$ and t_i' in type space $\mathcal{T}'$ with $\tilde{\varphi}_i'(t_i') = \varphi_i'$,

$$E_i(t_i, \mathcal{T}, \mathcal{M}^*) \cap E_i(t_i', \mathcal{T}', \mathcal{M}^*) = \emptyset.$$

Conversely, payoff types φ_i and φ_i' are strategically indistinguishable if, for every mechanism $\mathcal{M}$, there exist types t_i in type space $\mathcal{T}$ with $\tilde{\varphi}_i(t_i) = \varphi_i$ and t_i' in type space $\mathcal{T}'$ with $\tilde{\varphi}_i'(t_i') = \varphi_i'$ such that

$$E_i(t_i, \mathcal{T}, \mathcal{M}) \cap E_i(t_i', \mathcal{T}', \mathcal{M}) \neq \emptyset.$$

Bergemann and Morris (2009) present a characterization of strategically indistinguishable payoff types and show that strong interdependence in utilities gives rise to strategic indistinguishability. For example, in a quasi-linear environment where agent i has payoff type $\varphi_i \in [0, 1]$ and his valuation of an object is given by $v_i(\varphi) = \varphi_i + \gamma \sum_{j \neq i} \varphi_j$ for some $\gamma \in \mathbb{R}_+$, two distinct payoff types of any agent are strategically distinguishable if and only if $\gamma \leq \frac{1}{I-1}$.

8.3 Strategic Revealed Preference

Suppose we knew that an agent i would choose a_1 when playing mechanism $\mathcal{M}_1$, a_2 when playing mechanism $\mathcal{M}_2$, and so on. This might be because the agent made these choices in real time (and we knew his/her preferences—and implicitly information—were stable over time), or these might reflect hypothetical choices that the agent would make. If we had a finite data set given by $(a_k, \mathcal{M}_k)_{k=1}^K$, we could ask if there exists a type that could have generated that set of data by rational strategic choice. If we interpret rational strategic choice as choosing according to some solution concept, say, IPCR, i.e., then this "strategic revealed preference" question becomes: is there a type t_i in some type space $\mathcal{T}$ such that $a_k \in R_i(t_i, \mathcal{T}, \mathcal{M}_k)$ for every k?

This is a strategic analogue to the classical revealed preference question of Afriat (1967). In the single person case, without the linear indifference curves generated by expected utility preferences over lotteries, we know that a finite data set is consistent with a rational preference if and only if it satisfies the weak axiom of revealed preference (WARP). To get to our strategic revealed preference question described above, we must first require the outcome space to be a lottery space and impose expected utility preferences, which will require independence as well as WARP in the data. Second, we must translate a choice problem, where an agent picks a most preferred outcome from a set of lotteries, to a strategy setting where many agents make simultaneous (and perhaps interdependent) choices. Our mechanism is a many agent choice problem where outcomes depend not only on an agent's choice but also on others' choices.

Our characterization of strategic distinguishability answers a related but different question. Suppose that all the data that you have observed so far are consistent with an agent being type t_i

or type t_i'. Does there exist a mechanism by which one could be sure to distinguish them at the next round? It would be a natural next step to ask how much distinguishing could be done with smaller mechanisms and thus give a characterization of behavioral implications of interdependent preferences in a small set of mechanisms rather than quantify over all mechanisms.

There is a small existing literature developing strategic analogues of classic single agent decision theory. Sprumont (2000) considers static Nash equilibrium in static games, and thus may be the closest to our setting. But the extension from one agent to many agent choice problems is carried out in a very different way. First, he does not consider mixed strategies and does not maintain—as we do—the hypothesis of expected utility preferences. Second, and more importantly, our many agent decision problems (mechanisms) put no structure on the set of choices—there may be arbitrary action sets—but the outcome function may impose restriction. For example, the outcome resulting from one action profile may be identical to that resulting from another action profile, and we implicitly assume that there is common certainty of this fact. By contrast, Sprumont (2000) fixes agents' finite action sets and studies choices when there is common certainty that they are restricted to subsets of these actions sets. But he imposes no restrictions on how the outcomes from different action profiles may relate to each other.

A Calculations for the Example of Section 3.2

Observe that on the "reduced" complete information type space (without redundant types), agent 1 must choose out U in equilibrium. If agent 1 is sure her opponent is choosing L, her payoff gain to choosing U is

$$\frac{1}{2} + \frac{1}{2}\left(\frac{1}{3}\right) - \frac{151}{240} = \frac{160 - 151}{240} = \frac{3}{80} > 0;$$

but if an agent is sure her opponent is choosing R, her payoff gain to choosing out is

$$\frac{4}{3}\left(\frac{1}{3}\right) - \frac{4}{3}\left(\frac{1}{3} - \varepsilon\right) = \frac{4}{3}\varepsilon > 0.$$

On the "rich" Harsanyi type space (with redundant types), there will also be an equilibrium where all types choose (U, L) for sure. Thus types with the same preference hierarchy do indeed have an equilibrium action in common, as shown by our main theorem. However, there will also a strict equilibrium where, for agent 1, type l chooses U and type h chooses D, and for agent 2, type l chooses L and type h chooses R. Under this strategy profile, when agent 1 is type l, her expected payoff to choosing U is

$$\frac{5}{9}\left(\frac{1}{2}\left(1 + \frac{4}{15}\right)\right) + \frac{4}{9}\left(\frac{1}{3}\left(1 + \frac{5}{12}\right)\right);$$

while her expected payoff to choosing D is

$$\frac{5}{9}\left(\frac{151}{240}\right) + \frac{4}{9}\left(\left(\frac{1}{3} - \varepsilon\right)\left(1 + \frac{5}{12}\right)\right);$$

This gain to choosing U is then

$$\frac{5}{9}\left(\frac{152}{240} - \frac{151}{240}\right) + \frac{4}{9}\varepsilon\left(1 + \frac{5}{12}\right) > 0.$$

When agent 1 is type h, her expected payoff to choosing U is

$$\frac{4}{9}\left(\frac{1}{2}\left(1 + \frac{1}{4}\right)\right) + \frac{5}{9}\left(\frac{1}{3}\left(1 + \frac{2}{5}\right)\right);$$

while her expected payoff to choosing D is

$$\frac{4}{9}\left(\frac{151}{240}\right) + \frac{5}{9}\left(\left(\frac{1}{3} - \varepsilon\right)\left(1 + \frac{2}{5}\right)\right);$$

This gain to choosing D is then

$$\frac{4}{9}\left(\frac{151}{240} - \frac{150}{240}\right) - \frac{5}{9}\varepsilon\left(1 + \frac{2}{5}\right) > 0.$$

Under this strategy profile, when agent 2 is type l, his expected payoff to choosing L is

$$\frac{5}{9}\left(\frac{1}{2}\left(1 + \frac{4}{15}\right)\right) + \frac{4}{9}\left(\frac{151}{240}\left(\frac{5}{12}\right)\right) = \frac{379}{648};$$

35

while his expected payoff to choosing R is

$$\frac{5}{9}\left(\frac{1}{3}\left(1+\frac{4}{15}\right)\right) + \frac{4}{9}\left(\left(\frac{1}{3}-\varepsilon\right)\left(1+\frac{5}{12}\right)\right) = \frac{4}{9} - \frac{17}{27}\varepsilon;$$

This gain to choosing L is then

$$\frac{379 - 288 + 408\varepsilon}{648} = \frac{91 + 408\varepsilon}{648} > 0.$$

When agent 2 is type h, his expected payoff to choosing L is

$$\frac{4}{9}\left(\frac{1}{2}\left(1+\frac{1}{4}\right)\right) + \frac{5}{9}\left(\frac{151}{240}\left(\frac{2}{5}\right)\right) = \frac{451}{1080};$$

while his expected payoff to choosing R is

$$\frac{4}{9}\left(\frac{1}{3}\left(1+\frac{1}{4}\right)\right) + \frac{5}{9}\left(\left(\frac{1}{3}-\varepsilon\right)\left(1+\frac{2}{5}\right)\right) = \frac{4-2\varepsilon}{9};$$

This gain to choosing R is then

$$\frac{480 - 240\varepsilon - 451}{1080} = \frac{29 - 240\varepsilon}{1080} > 0.$$

B   Proof of Lemma 2

Suppose not. Then, there exist $\varepsilon > 0$ such that, for every $n \in \mathbb{N}$, there exist $\succsim_n, a_n, b_n \in P_w(X)$, measurable space C_n, perturbed outcome function $g_n \colon X \times P_w(X) \times C_n \to \Delta(Z)$ with $|g_n(\cdot, \cdot, c) - g^0| \le 1/n$ for every $c \in C_n$, and $\succsim'_n \in P_w(X \times X \times C_n)$ such that $d_P(\succsim_n, a_n) \le 1/n$, $d_P(\succsim_n, b_n) \ge \varepsilon$, $\mathrm{mrg}_1 \succsim'_n = \succsim_n$, $D^{1/n} \times C_n$ is $\succsim'_n$-certain, and $\mathrm{mrg}_{2,3} \succsim'_n$ weakly prefers $g_n(\cdot, b_n, \cdot)$ to $g_n(\cdot, a_n, \cdot)$. For each n, let $\nu_n \in \Delta(X \times X \times C_n \times (Z \setminus \{w\}))$ be a probability measure that represents $\succsim'_n$. Note that $\mu_n := \mathrm{mrg}_{1,4}\nu_n$ represents $\succsim_n$, and $\nu_n(D^{1/n} \times C_n \times (Z \setminus \{w\})) = 1$.[7] Since X is a compact metric space, by taking a subsequence if necessary, we can find $\succsim^*, b^* \in P_w(X)$ and $\mu^* \in \Delta(X \times (Z \setminus \{w\}))$ such that $\succsim_n \to \succsim^*$, $b_n \to b^*$, and $\mu_n \to \mu^*$ as $n \to \infty$. Note that $a_n \to \succsim^*$ as $n \to \infty$, $\succsim^* \ne b^*$, and μ^* represents $\succsim^*$.

Claim 2 *For every $k_0 \in \mathbb{N}$, there exists $n_0 \in \mathbb{N}$ such that, for every $n \ge n_0$ and $k, l \le k_0$, if $\succsim_n$ strictly prefers f_k to f_l, then a_n weakly prefers f_k to f_l.*

Proof. Fix any k_0. Suppose not. Then there exists a pair of $k, l \le k_0$ and a subsequence of $(\succsim_n, a_n)$ such that $\succsim_n$ strictly prefers f_k to f_l, and a_n strictly prefers f_l to f_k. Since $\succsim_n$ and a_n converge to the same limit, this is a contradiction. ∎

[7]$\mathrm{mrg}_\Lambda \nu_n$ with $\Lambda \subset \{1, 2, 3, 4\}$ denotes the marginal of ν_n with respect to the coordinates in Λ.

Claim 3 *There exist* k^*, l^* *such that* $\succsim^*$ *strictly prefers* f_{k^*} *to* f_{l^*} *while* b^* *strictly prefers* f_{l^*} *to* f_{k^*}.

Proof of Claim 3. Since $\succsim^* \neq b^*$, there exist $f, f' \in F_c(X)$ such that $\succsim^*$ and b^* have different preferences between f and f'. Since $\succsim^*$ and b^* satisfy the continuity, we can assume without loss of generality that $\succsim^*$ strictly prefers f to f' and b^* strictly prefers f' to f. (To see this, suppose, for example, that $\succsim^*$ is indifferent between f and f' while b^* strictly prefers f' to f. Then, replace f by $(1-\lambda)f + \lambda f''$ and f' by $(1-\lambda)f' + \lambda f'''$ for sufficiently small $\lambda > 0$, where $\succsim^*$ strictly prefers f'' to f'''. A similar trick works when $\succsim^*$ strictly prefers f to f' while b^* is indifferent between f to f'.) Since F is dense in $F_c(X)$ in the sup norm, by the continuity of $\succsim^*$ and b^*, we can assume $f, f' \in F$ without loss of generality. ∎

Claim 4 *There exists* $n_0 \in \mathbb{N}$ *such that, for every* $n \geq n_0$, b_n *strictly prefers* f_{l^*} *to* f_{k^*}.

Proof of Claim 4. Follows from $b_n \to b^*$ as $n \to \infty$. ∎

It follows from Claim 3 that there exists $\eta > 0$ such that

$$7\eta < 2^{-k^*-l^*+1} \int (f_{k^*} - f_{l^*}) d\mu^*.$$

Pick $k_0 \geq \max\{k^*, l^*\}$ such that

$$\sum_{\max\{k,l\}>k_0} 2^{-k-l+1} < \eta.$$

Claim 5 *There exists* $n_1 \in \mathbb{N}$ *such that, for every* $n \geq n_1$ *and* $k, l \in \mathbb{N}$ *such that* $\max\{k, l\} \leq k_0$, *if* $\succsim^*$ *strictly prefers* f_k *to* f_l, *then* a_n *also strictly prefers* f_k *to* f_l.

Proof of Claim 5. Follows from $a_n \to \succsim^*$ as $n \to \infty$. ∎

Note that

$$\left(\chi_{a_n}(f_k, f_l) - \chi_{b_n}(f_k, f_l)\right) \int f_k d\mu^* + \left(\chi_{a_n}(f_l, f_k) - \chi_{b_n}(f_l, f_k)\right) \int f_l d\mu^*$$

$$= \left(\chi_{a_n}(f_k, f_l) - \chi_{b_n}(f_k, f_l)\right) \int (f_k - f_l) d\mu^*$$

since $\chi_{a_n}(f_l, f_k) = 1 - \chi_{a_n}(f_k, f_l)$ and $\chi_{b_n}(f_k, f_l) = 1 - \chi_{b_n}(f_l, f_k)$.

Claim 6 *For every* $n \geq \max\{n_0, n_1\}$, *we have*

$$\left(\chi_{a_n}(f_k, f_l) - \chi_{b_n}(f_k, f_l)\right) \int (f_k - f_l) d\mu^* \begin{cases} = \int (f_{k^*} - f_{l^*}) d\mu^* & \text{if } (k, l) = (k^*, l^*), \\ \geq 0 & \text{if } \max\{k, l\} \leq k_0. \end{cases}$$

37

Proof of Claim 6. By Claims 4 and 5, $\chi_{a_n}(f_{k^*}, f_{l^*}) = 1$ and $\chi_{b_n}(f_{k^*}, f_{l^*}) = 0$; $\chi_{a_n}(f_k, f_l) = 1 \geq \chi_{b_n}(f_k, f_l)$ and $\int (f_k - f_l)d\mu^* > 0$ if $\succsim^*$ strictly prefers f_k to f_l; $\chi_{a_n}(f_k, f_l) = 0 \leq \chi_{b_n}(f_k, f_l)$ and $\int (f_k - f_l)d\mu^* < 0$ if $\succsim^*$ strictly prefers f_l to f_k; $\int (f_k - f_l)d\mu^* = 0$ if $\succsim^*$ is indifferent between f_k and f_l. ■

Claim 7 *There exists $n_2 \in \mathbb{N}$ such that, for every $n \geq n_2$ and $k \leq k_0$, we have*

$$\left| \int f_k d(\mathrm{mrg}_{2,4}\nu_n) - \int f_k d\mu_n \right| \leq \eta.$$

Proof of Claim 7. Since X is a compact metric space, every continuous function is uniformly continuous. Therefore, there exists $n_2 \in \mathbb{N}$ such that $|f_k(x) - f_k(x')| \leq \eta$ for every $k \leq k_0$ and $(x, x') \in D^{1/n_2}$. For every $n \geq n_2$, we have

$$\left| \int f_k d(\mathrm{mrg}_{2,4}\nu_n) - \int f_k d\mu_n \right|$$
$$= \left| \int (f_k(x')(z) - f_k(x)(z))d(\mathrm{mrg}_{1,2,4}\nu_n)(x, x', z) \right|$$
$$\leq \int |f_k(x')(z) - f_k(x)(z)|d(\mathrm{mrg}_{1,2,4}\nu_n)(x, x', z) \leq \eta$$

since $|f_k(x')(z) - f_k(x)(z)| \leq \eta$ for $(\mathrm{mrg}_{1,2,4}\nu_n)$-almost every (x, x', z). ■

We can now provide the proof of Lemma 2.

Proof of Lemma 2. Since $\mu_n \to \mu^*$ as $n \to \infty$, there exists $n \geq \max\{n_0, n_1, n_2, 1/\eta\}$ such that, for every $k \leq k_0$, $|\int f_k d\mu_n - \int f_k d\mu^*| < \eta$. We decompose $\int (g_n(\cdot, a_n, \cdot) - g_n(\cdot, b_n, \cdot))d(\mathrm{mrg}_{2,3,4}\nu_n)$ into the following four terms:

$$\int (g_n(\cdot, a_n, \cdot) - g_n(\cdot, b_n, \cdot))d(\mathrm{mrg}_{2,3,4}\nu_n)$$
$$= \sum_{\max\{k,l\} \leq k_0} 2^{-k-l+1}(\chi_{a_n}(f_k, f_l) - \chi_{b_n}(f_k, f_l)) \int f_k d\mu^*$$
$$+ \sum_{\max\{k,l\} \leq k_0} 2^{-k-l+1}(\chi_{a_n}(f_k, f_l) - \chi_{b_n}(f_k, f_l)) \left(\int f_k d(\mathrm{mrg}_{2,4}d\nu_n) - \int f_k d\mu^* \right)$$
$$+ \sum_{\max\{k,l\} > k_0} 2^{-k-l+1}(\chi_{a_n}(f_k, f_l) - \chi_{b_n}(f_k, f_l)) \int f_k d(\mathrm{mrg}_{2,4}\nu_n)$$
$$+ \int [(g_n(\cdot, a_n, \cdot) - g^0(\cdot, a_n)) - (g_n(\cdot, b_n, \cdot) - g^0(\cdot, b_n))]d(\mathrm{mrg}_{2,3,4}\nu_n).$$

38

The first term is larger than 7η by Claim 6. The other terms are at least as large as -4η, $-\eta$, and -2η, respectively, since $\sum_{\max\{k,l\}\leq k_0} 2^{-k-l+1} < 2$, $|\chi_{a_n} - \chi_{b_n}| \leq 1$,

$$\left| \int f_k d(\mathrm{mrg}_{2,4} d\nu_n) - \int f_k d\mu^* \right|$$
$$\leq \left| \int f_k d(\mathrm{mrg}_{2,4} d\nu_n) - \int f_k d\mu_n \right| + \left| \int f_k d\mu_n - \int f_k d\mu^* \right|$$
$$< 2\eta$$

by Claim 7, $\sum_{\max\{k,l\}>k_0} 2^{-k-l+1} < \eta$, $|f_k| \leq 1$, and $|g_n(\cdot,\cdot,c) - g^0| \leq 1/n \leq \eta$ for every $c \in C_n$. Thus $\succsim_n'$ strictly prefers $g_n(\cdot, a_n, \cdot)$ to $g_n(\cdot, b_n, \cdot)$, which is a contradiction. ∎

C λ-Continuity

We present now a version of the universal preference type space where there is common certainty that all agents' preferences are λ-continuous for some fixed $\lambda > 0$. Such spaces include preference type spaces with the worst outcome property, studied in the body of this paper, and other settings, such as finite type spaces of Abreu and Matsushima (1992) and "compact and continuous" type spaces.

Let $P_0(X)$ be the set of binary relations over $F(X)$ that satisfy completeness, transitivity, independence, continuity and monotone continuity, and are not indifferent over lotteries. $P_0(X)$ is endowed with the σ-algebra generated by $\{\succsim \in P_\lambda(X) \mid f \succsim f'\}$ for any $f, f' \in F(X)$. When X is a compact metrizable space, then $P_0(X)$ is endowed with the topology generated by $\{\succsim \in P_\lambda(X) \mid f \succ f'\}$ for any $f, f' \in F_c(X)$. (Recall that $F_c(X) \subseteq F(X)$ is the set of continuous acts.)

For each $\lambda > 0$, let $P_\lambda(X)$ be the set of preferences in $P_0(X)$ that satisfies λ-continuity. For any pair of outcomes $z, z' \in Z$ with $z \neq z'$, let $P_{z,z'}(X)$ be the set of preferences in $P_0(X)$ such that $z \succsim z'' \succsim z'$ for any $z'' \in Z$. We have $P_0(X) = \bigcup_{\lambda>0} P_\lambda(X) = \bigcup_{z\neq z'} P_{z,z'}(X)$. For any pair of outcomes $z, z' \in Z$ with $z \neq z'$ and $\lambda > 0$, let $P_{z,z',\lambda}(X) = \succsim \in P_\lambda(X) \cap P_{z,z'}(X)$.

Let $M_{z,z'}(X \times Z)$ be the set of finite signed measures μ on $X \times Z$ such that $1 = \mu(X \times \{z\}) \geq \mu(X \times \{z''\}) \geq \mu(E \times \{z'\}) = 0$ for any $z'' \in Z$ and any measurable $E \subseteq X$. A preference relation $\succsim$ belongs to $P_{z,z'}(X)$ if and only if there exists a unique $\mu \in M_{z,z'}(X \times Z)$ such that

$$f \succsim f' \Leftrightarrow \int_{X\times Z} f(x)(z)d\mu(x,z) \geq \int_{X\times Z} f'(x)(z)d\mu(x,z)$$

for any $f, f' \in F(X)$. Thus, we can identify $P_{z,z'}(X)$ and $M_{z,z'}(X \times Z)$ (endowed with the weak* topology if X is compact and metrizable). Moreover, if $\succsim \in P_{z,z'}(X)$ is λ-continuous, then $\|\mu\| \leq |Z|(1-\lambda)/\lambda$; if $\|\mu\| \leq (1-\lambda)/\lambda$, then $\succsim$ is λ-continuous. Let $M_{z,z',r}(X \times Z) = \{\mu \in M_{z,z'}(X \times Z) \mid \|\mu\| \leq r\}$.

For any signed measure μ on $X \times Z$, let $|\mu|$ denote the total variation measure on $X \times Z$, i.e., $|\mu|(E) = \|\mu(\cdot \cap E)\|$ for each measurable $E \subseteq X \times Z$.

C.1 Compactness and Metrizability

Proposition 5 *If X is compact and metrizable and $\lambda > 0$, then $P_\lambda(X)$ is also compact and metrizable.*

Proposition 5 follows from the next two lemmas.

Lemma 3 *If X is compact and metrizable, then $P_0(X)$ is Hausdorff.*

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'. By the trick we used in the proof of Claim 3 in Appendix B, we can assume without loss of generality that $f \succ f'$ and $f' \succ' f$.

Let μ and μ' be finite signed measures on $X \times Z$ that represent $\succsim$ and $\succsim'$, respectively. Let $\nu = |\mu| + |\mu'|$. We define the L^1-norm on measurable functions $\varphi \colon X \times Z \to \mathbb{R}$ (after identifying all $|\mu|$-a.e. equivalent functions) by

$$\|\varphi\| = \int_{X \times Z} |\varphi(x, z)| d\nu(x, z).$$

Since $F_c(X)$ is norm dense in $F(X)$ (Aliprantis and Border (1999, Theorem 12.9)), we can assume that f and f' are continuous. ■

Lemma 4 *If X is compact and metrizable and $\lambda > 0$, then $P_{z,z',\lambda}(X)$ is compact and metrizable for $z, z' \in Z$ with $z \neq z'$.*

Proof. Recall that $P_{z,z',\lambda}(X) \subset M_{z,z',r}(X \times Z)$ with $r = |Z|(1 - \lambda)/\lambda$. By the Riesz representation theorem and Alaoglu's theorem, $M_{z,z',r}(X \times Z)$ is weak*-compact. Also, by the Stone-Weierstrass theorem and Aliprantis and Border (1999, Theorem 6.34), $M_{z,z',r}(X \times Z)$ is weak*-metrizable. Thus we only need to show that $P_{z,z',\lambda}(X)$ is a closed subset of $M_{z,z',r}(X \times Z)$.

Take any sequence $\{\succsim_n\}$ on $P_{z,z',\lambda}(X)$ that converges to $\succsim$. We want to show that

$$(1 - \lambda)z + \lambda f \succsim (1 - \lambda)z' + \lambda f' \tag{3}$$

for any $f, f' \in F(X)$. Note that (3) for continuous acts $f, f' \in F_c(X)$ immediately follows from the definition of the topology on $P_0(X)$.

Pick a signed measure $\mu \in M_{z,z',r}(X \times Z)$ that represents $\succsim$. As in the proof of Lemma 3, $F_c(X)$ is dense in $F(X)$ with respect to the L^1-norm with measure $|\mu|$. Thus (3) extends from $F_c(X)$ to $F(X)$. ∎

Note that Lemma 5 relies on λ-continuity. To see this, notice that the set of finite signed measures (without a uniform bound on total variations) on an infinite metric space is neither weak*-compact nor metrizable.

Proof of Proposition 5. By Lemma 4, $P_\lambda(X)$ is a finite union of compact subspaces $P_{z,z',\lambda}(X)$, which is compact. Also, by Lemmas 3 and 4, $P_\lambda(X)$ is a finite union of closed and metrizable subspaces $P_{z,z',\lambda}(X)$, which is metrizable. (See Nagata (1985, Theorem 6.12), which follows from the Nagata-Smirnov metrization theorem.) ∎

C.2   Preference Type Spaces

We define a preference type space as $\mathcal{T} = (T_i, \pi_i)_{i \in \mathcal{I}}$, where, for each $i \in \mathcal{I}$, T_i is a measurable space of agent i's types, and $\pi_i \colon T_i \to P_0(\Theta \times T_{-i})$ is a measurable mapping that maps his types to preferences.

As we argued in Section 8.1, λ-continuity is a weak requirement. There is common certainty of λ-continuity with some $\lambda > 0$ if there is common certainty of the worst outcome property, or if the preference type space is finite. The second sufficient condition is generalized to compact and continuous type spaces as follows. We say that a preference type space $\mathcal{T} = (T_i, \pi_i)_{i \in \mathcal{I}}$ is *compact and continuous* if, for each $i \in \mathcal{I}$, T_i is compact and metrizable, and $\pi_i \colon T_i \to P_0(\Theta \times T_{-i})$ is continuous.

Proposition 6 *If a preference type space is compact and continuous, then there exists $\lambda > 0$ such that there is common certainty that preferences are λ-continuous.*

This follows immediately from the following lemma.

Lemma 5 *Assume that X is a compact metric space. For any compact subset Q of $P_0(X)$, there exists $\lambda > 0$ such that any preference in Q is λ-continuous.*

Proof. For each pair of outcomes $z, z' \in Z$ with $z \neq z'$, let $Q_{z,z'} = Q \cap P_{z,z'}(X)$. Since $P_{z,z'}(X)$ is closed in $P_0(X)$, $Q_{z,z'}$ is a weak*-compact subset of $M_{z,z'}(X \times Z)$. By the uniform boundedness principle, $Q_{z,z'}$ is bounded in the total variation norm. Pick $r_{z,z'} < \infty$ such that $Q_{z,z'} \subseteq M_{z,z',r_{z,z'}}(X \times Z)$. Then $Q_{z,z'} \subseteq P_{z,z',\lambda_{z,z'}}$ with $\lambda_{z,z'} = 1/(1 + r_{z,z'})$. Then any preference in Q is λ-continuous with $\lambda = \min_{z \neq z'} \lambda_{z,z'}$. ∎

C.3 The Universal Preference Type Space

The construction of the universal preference type space is analogous to Section 4.3. Given Proposition 5, all we have to do is to modify probability measures to signed measures. We use a generalization of the Kolmogorov extension theorem given by Van Haagen (1981), which requires coherency as well as uniform boundedness of total variations.

More specifically, let $X_{\lambda,0} = \{*\}$ and $X_{\lambda,n} = X_{\lambda,n-1} \times P_\lambda(\Theta \times X_{\lambda,n-1}^{I-1})$ for each $n \geq 1$. Let $X_{\lambda,\infty} = \prod_{n=0}^\infty P_\lambda(\Theta \times X_{\lambda,n}^{I-1})$. For a pair $z, z' \in Z$ with $z \neq z'$, let $Y_{z,z',\lambda,0} = \prod_{n=0}^\infty M_{z,z',|Z|(1-\lambda)/\lambda}(\Theta \times X_{\lambda,n}^{I-1} \times Z)$. Note that, for any $\{\mu_n\} \in Y_{z,z',\lambda,0}$, $\{\mu_n\}$ has uniformly bounded total variations. Let $Y_{z,z',\lambda,1}$ be the set of coherent hierarchies of signed measures in $Y_{z,z',\lambda,0}$, i.e., $\{\mu_n\} \in Y_{z,z',\lambda,0}$ such that $\mathrm{mrg}_{\Theta \times X_{\lambda,n-2}^{I-1} \times Z} \mu_n = \mu_{n-1}$ for any $n \geq 2$.

For each $\{\mu_n\} \in Y_{z,z',\lambda,1}$, by Van Haagen's (1981) generalization of the Kolmogorov extension theorem, there exists a signed measure μ_∞ on $\Theta \times X_{\lambda,\infty}^{I-1} \times Z$ such that $\mathrm{mrg}_{\Theta \times X_{\lambda,n-1}^{I-1} \times Z} \mu_\infty = \mu_n$ for any $n \geq 1$ and $\|\mu_\infty\| = \sup_n \|\mu_n\| \leq |Z|(1-\lambda)/\lambda$. It is easy to check that $\mu_\infty(\Theta \times X_{\lambda,\infty}^{I-1} \times \{z\}) = 1$ and $\mu_\infty(\Theta \times X_{\lambda,\infty}^{I-1} \times \{z'\}) = 0$, so we have $\mu_\infty \in M_{z,z',|Z|(1-\lambda)/\lambda}(\Theta \times X_{\lambda,\infty}^{I-1} \times Z)$. Thus we construct a homeomorphism $\psi_{z,z',\lambda} \colon Y_{z,z',\lambda,1} \to M_{z,z',|Z|(1-\lambda)/\lambda}(\Theta \times X_{\lambda,\infty}^{I-1} \times Z)$. Let $T_{\lambda,1}$ be the set of all coherent hierarchies of preferences in $X_{\lambda,\infty}$, i.e., $\{\succsim_n\} \in X_{\lambda,\infty}$ such that $\mathrm{mrg}_{\Theta \times X_{\lambda,n-2}^{I-1}} \succsim_n = \succsim_{n-1}$ for any $n \geq 2$. We convert $\psi_{z,z',\lambda}$ to a mapping between preference spaces and obtain a homeomorphism $\psi_{\lambda,P} \colon T_{\lambda,1} \to P_\lambda(\Theta \times X_{\lambda,\infty}^{I-1})$.

For $n \geq 2$, let
$$T_{\lambda,n} = \{t \in T_{\lambda,1} \mid \Theta \times T_{\lambda,n-1}^{I-1} \text{ is } \psi_{\lambda,P}(t)\text{-certain}\}.$$

For $i \in \mathcal{I}$, let $T_{i,\lambda}^* = \bigcap_{n=1}^\infty T_{\lambda,n}$ and a homeomorphism $\pi_{i,\lambda}^* = \psi_{\lambda,P}|_{T_{i,\lambda}^*} \colon T_{i,\lambda}^* \to P_\lambda(\Theta \times T_{-i,\lambda}^*)$. Thus we obtain $\mathcal{T}_\lambda^* = (T_{i,\lambda}^*, \pi_{i,\lambda}^*)_{i \in \mathcal{I}}$, the universal preference type space in which there is common certainty of λ-continuity.

C.4 Strategic Distinguishability

Once we understand equivalent preference hierarchies and interim preference correlated rationalizability (IPCR) as notions that respect λ-continuity, Theorems 1 and 2 hold verbatim for every $\lambda > 0$. Proposition 4 also holds, but now the construction of truth-telling mechanism $\mathcal{M}^*$ depends both on ε and λ. Proofs require minor modifications. For example, in the proof of Claim 7 in Appendix B, since ν_n is no longer a probability measure in $\Delta(X \times X \times W \times (Z \setminus \{w\}))$, but a signed measure in $M_{z,z',|Z|(1-\lambda)/\lambda}(X \times X \times W \times Z)$ with some $z, z' \in Z$, the last inequality needs to be replaced by
$$\int |f_k(x')(z) - f_k(x)(z)| d(\mathrm{mrg}_{1,2,4} \nu_n)(x, x', z) \leq \eta \|\nu_n\| \leq \eta \frac{|Z|(1-\lambda)}{\lambda}.$$

References

[1] D. Abreu and H. Matsushima (1992), "Virtual Implementation in Iteratively Undominated Actions," at http://www.princeton.edu/~dabreu/index_files/virtual%20implementation-incomplete.pdf

[2] S. Afriat, "The Construction of a Utility Function from Expenditure Data," *International Economic Review* 8, 66–77.

[3] C. D. Aliprantis and K. C. Border (1999), *Infinite Dimensional Analysis: Hitchhiker's Guide*, second edition. Springer.

[4] F. J. Anscombe and R. J. Aumann (1963), "A Definition of Subjective Probability," *Annals of Mathematical Statistics* 34, 199–205.

[5] P. Battigalli and M. Siniscalchi (2003) "Rationalization and Incomplete Information," *Advances in Theoretical Economics* 3 (1), Article 3.

[6] D. Bergemann and S. Morris (2009). "Virtual Robust Implementation," *Theoretical Economics* 4, 45–88.

[7] A. Brandenburger and E. Dekel (1993), "Hierarchies of Beliefs and Common Knowledge," *Journal of Economic Theory* 59, 189–198.

[8] E. Dekel, D. Fudenberg, and S. Morris (2006), "Topologies on Types," *Theoretical Economics* 1, 275–309.

[9] E. Dekel, D. Fudenberg, and S. Morris (2007), "Interim Correlated Rationalizability," *Theoretical Economics* 2, 15–40.

[10] A. Di Tillio (2008), "Subjective Expected Utility in Games," *Theoretical Economics* 3, 287–323.

[11] J. C. Ely and M. Pęski (2006), "Hierarchies of Belief and Interim Rationalizability," *Theoretical Economics* 1, 19–65.

[12] L. G. Epstein and T. Wang (1996), ""Beliefs about Beliefs" without Probabilities," *Econometrica* 64, 1343–1373.

[13] F. Gul and W. Pesendorfer (2007), "The Canonical Space for Behavioral Types," mimeo.

[14] J. O. Ledyard (1986), "The Scope of the Hypothesis of Bayesian Equilibrium," *Journal of Economic Theory* 39 (1), 59–82.

[15] D. K. Levine (1998), "Modeling Altruism and Spitefulness in Experiments," *Review of Economic Dynamics* 1, 593-622.

[16] Q. Liu (2009), "On Redundant Types and Bayesian Formulation of Incomplete Information," *Journal of Economic Theory*, forthcoming.

[17] J.-F. Mertens and S. Zamir (1985), "Formulation of Bayesian Analysis for Games with Incomplete Information," *International Journal of Game Theory* 14 (1), 1–29.

[18] P. Milgrom (2004), *Putting Auction Theory to Work*. Cambridge, England: Cambridge University Press.

[19] S. Morris and S. Takahashi (2010), "Common Certainty of Rationality," in progress.

[20] C. Müller (2009), "Robust Virtual Implementation under Common Strong Belief in Rationality".

[21] R. Myerson (1991), *Game Theory: Analysis of Conflict*. Cambridge, MA: Harvard University Press

[22] J. Nagata (1985), *Modern General Topology*, second revised edition. North-Holland.

[23] H. Paarsch and H. Hong (2006), *An Introduction to the Structural Econometrics of Auction Data*. Cambridge, MA: M.I.T. Press.

[24] A. Penta (2009), "Robust Dynamic Mechanism Design."

[25] T. Sadzik (2010), "Beliefs Revealed in Bayesian-Nash Equilibrium," mimeo.

[26] Y. Sprumont (2000), "On the Testable Implications of Collective Choice Theories," *Journal of Economic Theory* 93, 205–232.

[27] A. J. Van Haagen (1981), "Finite signed measures on function spaces," *Pacific Journal of Mathematics* 95 (2), 467–482.

[28] J. Weibull (2004), "Testing Game Theory," in *Advances in Understanding Strategic Behavior; Game Theory, Experiments and Bounded Rationality. Essays in Honour of Werner Guth*. Edited by Steffen Huck. Palgrave Macmillan.