REPUTATION WITHOUT COMMITMENT

JONATHAN WEINSTEIN AND MUHAMET YILDIZ

Abstract. In the reputation literature, players have commitment types which represent the possibility that they do not have standard payoffs but instead are constrained to follow a particular plan. In this paper, we show that arbitrary commitment types can emerge from incomplete information about the stage payoffs. In particular, any finitely repeated game with commitment types is strategically equivalent to a standard finitely repeated game with incomplete information about the stage payoffs, such that the types with identical solutions have almost identical prior probability in two games. Then, classic reputation results can be achieved with uncertainty concerning only the stage payoffs.

JEL Numbers: C72, C73.

1. Introduction

The reputation literature relies on the existence of commitment types. These types are not strategic but are certain to follow a particular plan. Since the seminal work of Kreps, Milgrom, Roberts, and Wilson (1982) (henceforth, Gang of Four), it has been well-established that inclusion of commitment types may alter the solution dramatically, as this may entice the original “rational” types to imitate the commitment types, in order to form a reputation for playing according to the committed plan. Building on this insight, a large literature has emerged, with applications in a wide range of areas.

Of course, commitment types can be modeled by using a payoff function that rewards a player who follows a specific plan. Nevertheless, such payoff functions often contradict the underlying structure of the original game. For example, the tit-for-tat types used by Gang of Four in the analysis of finitely repeated prisoners dilemma cannot be justified by a payoff function that is consistent with a repeated game, i.e. a (possibly discounted) sum of stage-game payoffs. The only commitment types that arise within such an additive structure are

Date: Preliminary Version: August 2012.

1We refer to the textbook of Mailath and Samuelson (2006) for a review.
those who commit to playing the same action throughout the game, and these would not have any impact on behavior in a repeated prisoners dilemma game.

The justification for commitment types is important for the interpretation of the results. If one can derive an outcome (say cooperation in the finite prisoner’s dilemma) using only types that arise from a natural payoff structure (such as a repeated game with different stage payoffs), one can then interpret the result as an outcome of incomplete information within a rational framework. On the other hand, some payoff functions, such as one that rewards a player only if he follows tit-for-tat, seem irrational in a repeated-game context. They may instead reflect psychological anomalies and super-game concerns (motivations that lie outside the game), such as maintaining reputation in a broader context. If one needs to introduce such payoff functions, then the result is probably best seen as reflecting the issues relating to the super-game concerns, psychological anomalies or irrationality that have been overlooked in the original model. (That may be why the commitment types are often referred to as crazy types and treated within the context of modeling irrationality, as in Kreps (1993).)

In this paper, we show that every commitment type can be obtained as a type with a unique rationalizable plan within the additive payoff structure of finitely repeated games. Moreover, every such game with commitment types is strategically equivalent to a standard repeated game with incomplete information about the stage-game payoff function, such that the types with identical solutions have almost identical prior probabilities. Therefore, one can justify any reputational model of finitely repeated game within a rationalistic model of incomplete information without violating the structural assumptions of repeated games. This is true provided that one allows sufficient variations in stage-game payoffs and considers a rich set of information structures.

Of course, one may further restrict the structure of the stage-game payoff functions. For example, one may not be willing to accept the possibility that cooperation dominates defection in the context of the prisoners’ dilemma game. In that case, one may not be able to justify some commitment types, as our benchmark above could be too permissive. We also prove an opposite benchmark, showing that one needs some amount of variations in the stage-game payoffs in order to have any reputational effect. We show that if the stage game is dominance-solvable and the stage game payoffs are restricted to be a sufficiently small
neighborhood of the original stage-game payoff function, then the unique sequential equilibrium of the repeated game with incomplete information prescribes all players to repeat the stage-game solution throughout the game (as in the subgame-perfect equilibrium of the complete information version), regardless of the length of the game.

2. Preview of Results

In this section, we preview our result more carefully on the example analyzed by Gang of Four: the finitely-repeated prisoner’s dilemma game in which player 1 may be committed to tit-for-tat with small probability.

Consider the repeated game in which the following prisoner’s dilemma is repeated \( t \) times:

\[
\begin{array}{c|cc}
 & \text{Cooperate} & \text{Defect} \\
\hline
\text{Cooperate} & 5, 5 & 0, 6 \\
\text{Defect} & 6, 0 & 1, 1 \\
\end{array}
\]

All the previous moves are publicly observable (perfect monitoring), and the payoff of a player in the repeated game is the sum of his payoffs in the stage game above. A simple application of backward induction in this game yields the play of (Defect, Defect) at every history. Indeed, it is well known that the only Nash equilibrium outcome is playing (Defect, Defect) at every period.

Gang of Four consider an incomplete information game \( G \) in which player 1 may be committed to playing tit-for-tat. Player 1 has two types, a “rational” type \( \tau^*_1 \), whose payoffs and available moves are as in the repeated-prisoners’ dilemma game above, and a committed type \( \tau^{T4T}_1 \) which can only play tit-for-tat. That is, the latter type must play cooperate in the first round and imitate the last move of player 2 in the subsequent periods. The prior probability of \( \tau^{T4T}_1 \) is some small \( \varepsilon > 0 \). Player 2 still has one type \( \tau^*_2 \), which is “rational” as in the original game. Gang of Four shows that in any sequential equilibrium of the new game each rational type \( \tau^*_1 \) must play Cooperate at all but few periods. (Interestingly, by varying the set of commitment types, one can generate a rich set of equilibrium behavior, obtaining a Folk Theorem for long games (Fudenberg and Maskin, 1986)).

As we mentioned in the introduction, one can replicate the above equilibrium behavior with payoff uncertainty by assigning the payoff function of \( \tau^{T4T}_1 \) as 1 at histories at which
player 1 plays according to tit-for-tat and 0 at all other histories. Here, the solution concept is sequential equilibrium with the restriction that player 2 assigns probability 1 on \( \tau_1 \) off the path. Such a payoff function is incompatible with the repeated game payoff structure, and one cannot replicate the commitment to tit-for-tat by simply modifying the stage-game payoff function for \( \tau_1^{T4T} \). Indeed, such modifications can lead to only two commitment types: the type that plays Cooperate throughout and the type that plays Defect throughout. Commitment to cooperation can be justified by the stage-game payoff function

<table>
<thead>
<tr>
<th></th>
<th>Cooperate</th>
<th>Defect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cooper</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Defect</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

for example. The inclusion of such simple commitment types cannot affect the behavior of rational types in this game, though in other games such as Cournot duopoly, it could result in a player becoming a Stackelberg leader by convincing the other player he is committed to a certain action.

Fortunately, the austere information structure above is not only structure we can consider. For any \( \epsilon' > \epsilon \), consider the games in which

1. all types can play all strategies and maximize the sum of stage-game payoffs, which need not be as in (PD);
2. the prior probability of the rational type profile \((\tau_1^*, \tau_2^*)\) is at least \(1 - \epsilon'\), and each \(\tau_i^*\) knows that his stage-game payoffs are as in (PD).

We construct such a game \(G'\) that is strategically equivalently to the game \(G\) in Gang of Four. The type space in \(G'\) contains types \(\tau_1^*, \tau_2^*\), a twin \(\bar{T}_1^{T4T}\) of the tit-for-tat type \(\tau_1^{T4T}\) in \(G\), and a number of other new types that we use to encode the beliefs of type \(\bar{T}_1^{T4T}\). Although \(\bar{T}_1^{T4T}\) is allowed to play any plan of action, he chooses to play tit-for-tat as his unique rationalizable plan. In \(G'\), the prior distribution puts probability \(1 - \epsilon'\) on \((\tau_1^*, \tau_2^*)\), \(\epsilon (1 - \epsilon') / (1 - \epsilon)\) on \((\bar{T}_1^{T4T}, \tau_2^*)\) and put the remaining small probability \((\epsilon' - \epsilon) / (1 - \epsilon)\) on the newly constructed types. As a result, rational type \(\tau_1^*\) is certain that he faces the rational type \(\tau_2^*\), and the rational type \(\tau_2^*\) in turn puts probability \(1 - \epsilon\) on \(\tau_1^*\) and probability \(\epsilon\) on the twin \(\bar{T}_1^{T4T}\) of \(\tau_1^{T4T}\). Hence, the strategic situation the rational types face is the same as in \(G\), except now \(\tau_2^*\) thinks that \(\bar{T}_1^{T4T}\) plays tit-for-tat as a result of some rational reasoning under incomplete information rather than as a result of commitment or an unconventional
payoff function. Therefore, under the broad set of solution concepts that are invariant to such changes, the solution sets for rational types \((\tau_1^*, \tau_2^*)\) are identical.

Three points are worth emphasizing. First, when \(\varepsilon\) and \(\varepsilon'\) are close, the probabilities of \((\tau_1^*, \tau_2^*)\) and \((\bar{t}_1^{T4T}, \tau_2^*)\) are approximately \(1 - \varepsilon\) and \(\varepsilon\), respectively. Hence, the type spaces of \(G\) and \(G'\) are nearly identical. Despite this, the twin \(\bar{t}_1^{T4T}\) has a unique rationalizable plan because \(\bar{t}_1^{T4T}\) believes that his own plan could have any non-negligible impact on his payoff only if he faces one of the newly constructed types. He finds these types unlikely, but they are likely enough to become his main concern. Second, the unique rationalizable plan emerges under some intricate beliefs that require a large number of new types for encoding, especially when the game is long. Nonetheless, we are able to encode such beliefs by putting only a negligible amount of prior probability on the new types. Third, our proof is based on typical reward and punishment mechanisms in repeated games as well as theories of social learning that rely on the existence of payoff functions as in (CC). One could object that such stage-games are too different from the original prisoners’ dilemma game to be conceivable.

Our second benchmark allows only small variations in the stage-game payoffs. Since Nash equilibrium is robust to such perturbations and (Defect, Defect) is the only Nash equilibrium here, one cannot expect to have any discontinuity in that front. This is similar to the fact that for a given \(\bar{t}\), (Defect, Defect) becomes the unique equilibrium outcome as \(\varepsilon \rightarrow 0\) in the Gang of Four example. Nevertheless, Gang of Four demonstrates that such continuity is misleading: for any given \(\varepsilon\), long cooperation is obtained when \(\bar{t}\) is large enough. In this paper, we prove something stronger: a uniform continuity result with respect to the stage game payoffs. According to our result, when the stage-game payoffs are sufficiently close to (PD) throughout the type space, (Defect, Defect) is the only sequential equilibrium outcome, no matter how large \(\bar{t}\) is.

Therefore, one needs to allow some substantial amount of variation in stage-game payoffs in order to provide an incomplete-information foundation for the commitment types. While the amount of necessary variation may depend on the details of the game and the commitment types at hand, our main result shows that one can always provide such a foundation as long as there is enough variation in allowable stage-game payoff functions.

Our results here build on our previous work on non-robustness in repeated games. In Weinstein and Yildiz (2012) we showed that, in any infinitely repeated game, any individually
rational and feasible outcome is the unique rationalizable outcome of an appropriately chosen perturbation which maintains common knowledge of the repeated-game structure and discounting criterion. A key lemma leading to this result showed that for any plan whatsoever, there is a type who follows this plan as a unique rationalizable action, although he believes in common knowledge of the repeated-game structure. An extension of this lemma to finitely repeated games plays an important role in our construction. Nonetheless, perturbations considered in the two work are very different: here we use the ex-ante perturbations that are used in the traditional reputation literature (and also by Kajii and Morris (1997) in the context of robustness), while in Weinstein and Yildiz (2012) we use the perturbations of interim beliefs in universal type space. The latter perturbations are known to be much more permissive, and the two approaches often yield quite different results.

We introduce the basic definitions and formulations in Section 3. In Section 4, we present our construction of a new type space in which the commitment types are replaced by types for which the committed action plan is uniquely rationalizable. In Section 5, we show that, for the original rational types, the constructed game is strategically equivalent to the model with commitment types, under a very broad set of solution concepts. After presenting our continuity result in Section 6, we conclude in Section 7. Some of the more complicated proofs are relegated to the Appendix.

3. Basic Definitions

We study, for simplicity, a standard 2-player finitely repeated game with perfect monitoring and normal-form stage games. We write $N = \{1, 2\}$ for the set of players, $T = \{0, 1, \ldots, \bar{t}\}$ for the set of dates $t$, and fix a finite set $A = A_1 \times A_2$ of stage-game strategy profiles $a = (a_1, a_2)$ in normal form. Note that, since we have perfect monitoring, the non-initial histories in the repeated game are of the form $h^t = (a^0, \ldots, a^{t-1})$ where $a^{t'}$ denotes the stage-game strategy profile played at date $t' \in T$. We write $h^0$ for the empty initial history, and write $H$ for the set of all non-terminal histories. An outcome path or a terminal history is a list $(a^0, \ldots, a^\bar{t})$; the set of all terminal histories is denoted by $Z$.

\footnote{Following the convention in game theory, we write $-i$ for the player $j \neq i$ and drop the subscript to denote the profiles, e.g., $x = (x_1, x_2) \in X = X_1 \times X_2$ and $X_{-1} = X_2$.}
The payoff vector from an outcome path \( (a^0, a^1, \ldots, a^t) \) in repeated game is simply the sum of the stage game payoffs:

\[
(3.1) \quad u(a^0, a^1, \ldots, a^t|g) = g(a^0) + g(a^1) + \cdots + g(a^t),
\]

where the function \( u = (u_1, u_2) \) denotes the payoffs from the repeated game and the function \( g = (g_1, g_2) \) denotes the payoffs from the stage game. While the particular stage payoffs are not necessarily known, this formula will be common knowledge throughout the games we study here. That is, it is common knowledge that the stage payoff function \( g \) is fixed throughout the game and that the players simply maximize the sum of these payoffs.

We write \( G^* = [0, 1]^N \times A \) for the set of all possible profiles \( g = (g_1, g_2) \) of stage-game payoff functions \( g_i : A \rightarrow [0, 1] \). Here, we put a uniform bound on the stage game payoffs so that small variations of the probability distributions on stage payoffs lead to small variations in expected payoffs, as in the reputation literature. This restriction strengthens our results.

We fix a complete-information repeated game in which it is common knowledge that the stage-game payoff function is a fixed \( g^* : A \rightarrow [0, 1]^2 \). This could, for example, be the repeated prisoner’s dilemma game. The payoff function of a rational type in the repeated game is given by the formula

\[
(3.2) \quad u^*_i(a^0, a^1, \ldots, a^t) = g^*_i(a^0) + g^*_i(a^1) + \cdots + g^*_i(a^t),
\]

and all action plans are available for rational types.

In the complete-information game, a strategy of a player \( i \) is a mapping \( s_i : h^t \mapsto s^t_i(h^t) \in A_i \), which maps each non-terminal history to strategy in the stage game. Because we analyze incomplete information games, however, we will avoid the word strategy for this mapping and call it instead an action plan, because we reserve the word strategy for mappings from types to action plans. The set of all action plans is denoted by \( S_i \). The outcome path induced by a profile \( (s_1, s_2) \) is denoted by \( z(s_1, s_2) \). We also allow (behavioral) mixed strategies and write \( \pm_i \) for the set of mixed action plans \( \sigma_i : h^t \mapsto \sigma^t_i(h^t) \in \Delta (A_i) \) for player \( i \).

We consider two kinds of elaboration, corresponding to two distinct ways in which the common-knowledge assumption in the complete information game may be relaxed. The first

\footnote{Discounting would not affect our results; setting the discount rate to 1 will make our proofs easier to follow.}
notion of relaxation is the one considered in the reputation literature—through commitment types.

**Definition 1.** An \( \varepsilon \)-elaboration with (one-sided) commitment types is a Bayesian game such that

- the sets of types for players 1 and 2 are \( \{ \tau_1^* \} \cup C \) and \( \{ \tau_2^* \} \), respectively, where \( C \subset S_1 \);
- the probability of \( \tau_1^* \) is \( 1 - \varepsilon \) and the probability of each \( c \in C \) is some \( \pi(c) \in [0, 1] \) that add up to \( \varepsilon \);
- the set of moves available to \( \tau_1^* \) is as in the repeated game above, while the only available action plan for type \( c \in C \) is \( c \),
- and the payoffs are as in the complete information game.

Such an elaboration can be denoted by a pair \( (C, \pi) \) of commitment types \( C \) and a vector of probabilities on \( C \).

Here, each action plan \( c \in C \) corresponds to a type of Player 1 who can only play \( c \). The incomplete information is only about whether Player 1 can play all action plans or has committed to a particular action plan. The type \( \tau_1^* \) that can play all plans is called the rational type while the types \( c \in C \), who can play only according to one plan of action, are called commitment types. The commitment types are also sometimes called crazy types.

A commitment type \( c \) is equivalent to a type with a payoff function that gives 1 if Player 1 plays according to \( c \) throughout and 0 if he ever deviates. Such a payoff function clearly violates the additive structure in (3.1). Indeed, except for the trivial commitment types who play the same stage game strategy throughout the repeated game, commitment types cannot be justified by a new payoff function that has the additive structure above.

We will next introduce type spaces in which all payoff functions satisfy the additive structure of (3.1). While they enforce this restriction on the repeated structure of the game, these spaces are also “richer” in that they may have a larger variety of types. Our type spaces will have two-sided incomplete information with a common prior. Formally, by a type space, we mean a list \( (\mathcal{G}, \mathcal{T}, \pi) \) where \( \mathcal{G} \subset \mathcal{G}^* \) is a finite set of payoff function profiles \( g \), \( \mathcal{T} = \mathcal{T}_1 \times \mathcal{T}_2 \) is the set of type profiles \( \tau = (\tau_1, \tau_2) \), and \( \pi \in \Delta(\mathcal{G} \times \mathcal{T}) \) is the common prior\(^4\) A Bayesian

\(^4\)Here, \( \Delta(X) \) denotes the set of all probability measures on the finite set \( X \).
**Repeated Game** (without commitment types) is a list \((N, A, (G, T, \pi))\). We should emphasize that this notation suppresses many important common-knowledge assumptions, such as the fact that the game is repeated, all previous actions are publicly observable (i.e., perfect monitoring), and the payoffs in the repeated game are given by the formula (3.1). A strategy of a player \(i\) in a Bayesian repeated game \((N, A, (G, T, \pi))\) is a mapping \(\sigma_i : T_i \rightarrow \Sigma_i\).

The second notion of elaboration allows richer type spaces and two-sided incomplete information, but does not allow any payoff function outside of the additive structure in (3.1):

**Definition 2.** An \(\varepsilon\)-elaboration without commitment types of a complete-information game \(g^*\) is a Bayesian game \((N, A, (G, T, \pi))\) with distinguished types \(\tau_1^*, \tau_2^*\) where

1. \((g^*, \tau^*) \in G \times T\),
2. \(\pi(g^*, \tau^*) = 1 - \varepsilon\), and
3. \(\pi(g_i^*|\tau_i^*) = 1\).

The first and second conditions state that the original complete information game is embedded in the elaboration and has a high ex-ante probability of \(1 - \varepsilon\). The last condition states that the rational types \((\tau_1^*, \tau_2^*)\) know their payoffs, and their payoffs are as in the original complete information game. The novelty in this definition is that the payoffs under all possible specifications in the Bayesian game satisfy the additive repeated game structure of (3.1). That is, the formula (3.1) remains common knowledge. In that sense, all the types in an elaboration without commitment types are rational, although we reserve the term rational for types \((\tau_1^*, \tau_2^*)\) as in the elaborations with commitment types.

Both elaborations above fall under the category of \(\varepsilon\)-elaborations as defined by Kajii and Morris (1997). An \(\varepsilon\)-elaboration without commitment types is a Kajii-Morris elaboration with the additional restriction that the formula (3.1) is common knowledge. While an \(\varepsilon\)-elaboration without commitment types presented in terms of uncertainty about the strategies, as discussed above, they could also be represented as Kajii-Morris elaborations with specific simple type space in which the formula (3.1) fails.

Finally, we review two standard concepts in game theory. First, for any Bayesian game, interim correlated rationalizability (henceforth ICR) is the outcome of iterated elimination of action plans for types that are never a weak best response, as defined by Dekel, Fudenberg,
and Morris (2007). We write \( S_i^\infty [\tau_i | G] \) for the set of all interim correlated rationalizable action plans for type \( \tau_i \in T_i \) in game \( G = (N, A, (G, T, \pi)) \). We will give a more detailed definition of ICR later in the construction. We just note here that ICR is the weakest known rationalizability concept for Bayesian games, and all the action plans that are played by a type with positive probability in any equilibrium are ICR for that type.

Second, we say that action plans \( s_i \) and \( s'_i \) are equivalent if \( z(s_i, s_{-i}) = z(s'_i, s_{-i}) \) for all action plans \( s_{-i} \in S_{-i} \), i.e., they lead to the same outcome no matter what strategy the other player plays. Note that \( s_i \) and \( s'_i \) are equivalent iff \( s_i(h^t) = s_i(h^t) \) for every history \( h^t \) in which \( i \) played according to \( s_i \) throughout; they may differ only in their prescriptions for histories that they preclude. Hence, in reduced form, action plans can be represented as mappings that maps the history of other players’ play into own stage game actions. We write \( \tilde{S}_i \) for the set of reduced-form action plans \( \tilde{s}_i \); these map each \((a_{-i}^t)_{0 \leq t < t}\) to some action \( a_i \in A_i \) in the stage game.

4. Irrelevance of Commitment Types

In this section, we show that one can replace each commitment type \( c \) with a regular type for which \( c \) is uniquely rationalizable with a slight perturbation of the prior distribution. Consequently, one can transform any elaboration with commitment types to an elaboration without commitment types, such that, from the point of view of rational types who believe in the ICR concept, the two elaborations are identical. Under ICR as well as a broader set of solution concepts, this will lead to the same set of solutions for the rational types in the two elaborations.

Proposition 1. For any \( \varepsilon' > \varepsilon > 0 \) and for any \( \varepsilon \)-elaboration \( G \) with commitment types \( (C, \pi) \) there exists a \( \varepsilon' \)-elaboration \( G' = (N, A, (G, T, \pi')) \) without commitment types in which the commitment types are replaced by types with unique rationalizable action plans:

1. \( \pi'(g^*, \tau_2^* | \tau_1^*) = 1 \) and \( \pi'(g^*, \tau_1^* | \tau_2^*) = \pi(g^*, \tau_1^* | \tau_2^*) = 1 - \varepsilon \), and

2. for every \( c \in C \) there exists \( \tau_1^c \in T_1 \) such that all ICR action plans of \( \tau_1^c \) are equivalent to \( c \), and \( \pi'(g^*, \tau_1^c | \tau_2^c) = \pi(g^*, c | \tau_2^c) = \pi(c) \).

Here, the first condition establishes that the interim beliefs of rational types regarding their own payoffs and “rationality” of their opponents are identical in the two elaborations. The
second condition establishes that each commitment type \( c \) is replaced by a type \( \tau_1^c \) for which following \( c \) is uniquely rationalizable, and that the rational type of player 2 in \( G' \) assigns the same probability to the type \( \tau_1^c \) as the rational type in \( G \) assigns to the commitment type \( c \).

The proposition thereby establishes that any \( \varepsilon \)-elaboration \( G \) with commitment types can be converted into an \( \varepsilon' \)-elaboration by replacing commitment types \( c \) with types \( \tau_1^c \) for whom \( c \) is the only rationalizable action plan in reduced form. These types follow \( c \) not because they are committed or have payoffs that are inconsistent with playing a repeated game but because their reasoning under their information leads them to do so. Moreover, from the point of view of the rational types these are the only types with positive probability, mirroring the elaboration with commitment types.

The equivalence is established despite the following constraints:

1. The repeated-game payoff structure is maintained throughout \( G' \). That is, it is common knowledge throughout that the payoff in the repeated game is the sum of the payoffs in the stage game, and that the stage game is fixed throughout the game. Type \( \tau_1^c \) knows all this and yet follows \( c \) as its unique rationalizable plan.
2. The ex-ante distribution \( \pi' \) in \( G' \) can be arbitrarily close to the distribution \( \pi \) in \( G \), in that \( \varepsilon' \) can be arbitrarily close to \( \varepsilon \).

Note that the commitment plans \( c \) can be arbitrarily complex. Hence, the types \( \tau_1^c \) that replace the commitment types \( c \) may hold highly complicated beliefs, and the elaboration without commitment types may contain a large number of new types that are used to encode these beliefs within the standard common prior type space. Despite this, the second condition requires that the ex-ante probability of these types are made arbitrarily small.

**Proof of Proposition 1.** The first step in our construction is the following lemma, which establishes that any given action plan \( s_i \) is the only rationalizable action for a type \( \tau_i^{s_i} \) from some common prior model. (The proof of Lemma 1 is the lengthiest step of the proposition and is relegated to the appendix.)

**Lemma 1.** For any \( s_i \in S_i \), there exists a Bayesian repeated game \( G^{s_i} = (N, A, (G^{s_i}, T^{s_i}, \pi^{s_i})) \) with a type \( \tau_i^{s_i} \in T_i^{s_i} \) such that

1. \( \pi^{s_i}(g, \tau) > 0 \) for every \( (g, \tau) \in G^{s_i} \times T^{s_i} \) and
(2) for every action plan $s'_i$, $s'_i \in S_i^\infty [\tau_i^{s_i} | G^{s_i}]$ if and only if $s'_i$ is equivalent to $s_i$.

By relabeling if necessary, we take all of the types above to be distinct from each other and from $\tau^*$, fixing also a unique type $\tau_i^{s_i}$ for each $s_i$. We construct $G' = (N, A, (G', T', \pi'))$ by setting

$$
G' = \{g^*, (0, g_2^*)\} \cup \bigcup_{c \in C} G^c$$

$$
T'_j = \{\tau_j^i\} \cup \bigcup_{c \in C} T_j^c \quad (\forall j \in N)
$$

$$
\pi'(g, \tau) = \begin{cases} 
1 - \varepsilon' & \text{if } (g, \tau) = (g^*, \tau^*) , \\
\frac{1 - \varepsilon'}{1 - \varepsilon} \pi(c) & \text{if } (g, \tau) = ((0, g_2^*), (\tau_1^i, \tau_1^*)), \\
\frac{\varepsilon' - \varepsilon}{(1 - \varepsilon)|C|} \pi^c(g, \tau) & \text{if } (g, \tau) \in G^c \times T^c, \\
0 & \text{otherwise},
\end{cases}
$$

where

$$
0(a) = 0 \quad (\forall a \in A).
$$

We now observe that $G'$ satisfies the properties in the proposition. Indeed, rational type $\tau_1^*$ of Player 1 assigns probability 1 on $(g^*, \tau_2^*)$. Likewise, since $\pi'(G' \times \{\tau_2^*\}) = (1 - \varepsilon') / (1 - \varepsilon)$, in the interim, $\tau_2^*$ assigns probability $1 - \varepsilon$ on $(g^*, \tau_1^*)$ and probability $\pi(c)$ on $\tau_1^i$ for each $c$. On the other hand, since the beliefs of type $\tau_1^i$ altered substantially when $(G^c, T^c, \pi^c)$ has been incorporated in $G'$, it is not clear that $\tau_1^i$ follows $c$ as the unique ICR action. The next lemma states that this is indeed the case.

**Lemma 2.** For any $c \in C$, $i \in N$, and any $\tau_i \in T_i^c$, $S_i^\infty [\tau_i | G'] = S_i^\infty [\tau_i | G^c]$; in particular, $S_i^\infty [\tau_i^c | G'] = c$.

This lemma completes the proof of the proposition; its proof is in the appendix. \(\square\)

Our proof has two main steps. The first, found in Lemma 1, is to construct a type space in which a given action plan is uniquely rationalizable for a type. We constructed such a type space in Weinstein and Yildiz (2012) for infinite-horizon repeated games, but without requiring that the constructed type space has a common prior, a property that is essential for our proposition here. In this paper, using the ideas in that construction, we first construct such a type space for finite-horizon games without common prior and then convert it to a
common-prior type space, using this time the ideas and the results developed by Lipman (2003) and Weinstein and Yildiz (2007).

The main economic ideas involved in these constructions come from social learning and reward and punishment mechanisms in repeated games. First, for the class of action plans that is consistent with learning in a single decision problem, we construct games in which players whose payoffs depend only on their own action update their optimal stage game strategies as they observe the past behavior of the other players and learn about their own payoff from the other players’ moves, as in social leaning. We show that such processes can generate any action plan that is consistent with individual learning as the unique ICR action plan. The generated action plans can be arbitrarily complex because individual learning can only put a relatively mild restriction on behavior, similar to the sure-thing principle. Accordingly, the type space generating these behavior can be highly complex, leading players to update their beliefs about own payoffs, other players’ beliefs, and other players’ beliefs about beliefs and so on, as they observe the other players’ moves. These ideas apply to the finite and infinite repeated games, and the proof is nearly identical for both cases. For infinitely repeated games, in Weinstein and Yildiz (2012) we extend the result to all action plans (including plans that contradicts the condition for individual learning), using a reward and punishment mechanism. Unfortunately, it is harder to come up with effective reward and punishment mechanisms for finite horizon games. After all, one cannot provide any future incentive in the last period. Hence, here, we use a more nuanced construction that combines social learning with a reward and punishment mechanism to extend the result to all action plans in finitely repeated games.

The second main step is to incorporate the above type spaces in one common prior model, replacing each commitment type with one of these type spaces. One must do this in such a way that (i) the original complete-information game still has high prior probability \( (1-\varepsilon') \), (ii) the interim beliefs of the rational types are as in the original elaboration with commitment types, and (iii) the types’ rationalizable behavior in the constructed type space remain the same after incorporating them into common prior model. The conditions (ii) and (iii) diametrically oppose each other, making the construction more difficult. To see this, note that (i) and (ii) require that the common prior \( \pi' \) puts a high probability on \( \tau'_i \), requiring that probability to be \( \frac{1-\varepsilon'}{1-\varepsilon} \pi(c) \) as in our proof. When \( \varepsilon \) and \( \varepsilon' \) are close, this probability is
approximately $\pi(c)$. When $\varepsilon$ and $\varepsilon'$ are close, this also requires that $\pi'$ puts a very small probability on $T^c$, the original type profiles in the constructed type space in the first step. That probability can be at most $(\varepsilon' - \varepsilon) / (1 - \varepsilon)$, which is negligible with respect to $\frac{1 - \varepsilon'}{1 - \varepsilon} \pi(c)$ when $\varepsilon$ and $\varepsilon'$ are close. These constraints make the belief of type $\tau_1^c$ in game $G'$ substantially different from the belief of the type $\tau_1^c$ in game $G^c$. In our construction, type $\tau_1^c$ in game $G'$ assigns probability

\begin{equation}
 p^c = \frac{|C| (1 - \varepsilon^c) \pi(c)}{|C| (1 - \varepsilon^c) \pi(c) + (\varepsilon' - \varepsilon) \pi^c(\tau)}
\end{equation}

on type $\tau_2^s$. Note that, for fixed $\pi(c)$, when $\varepsilon' - \varepsilon$ approaches 0, $p^c$ approaches 1. In contrast, $\tau_1^c$ in game $G^c$ assigns zero probability on $\tau_2^s$. Consequently, the belief hierarchies of the types in $G'$ can be quite different from the belief hierarchies of the types in $G^c$ with the same label, leading possibly to distinct set of ICR actions. We circumvent this problem with the following trick. We set the beliefs such that, whenever player 2 has type $\tau_2^s$, the payoff of type $\tau_1^c$ is 0 for every move in the stage game, making him indifferent among all outcomes. Since $p^c < 1$, his best responses are identical to his best responses conditional on the type of player 2 being other than $\tau_2^s$, thereby replicating the best responses of his twin in $G^c$. Since this was the only difference between the two type spaces, the rationalizable actions turn out to be identical in games $G^c$ and $G'$, as stated formally in Lemma 2.

Roughly speaking, from the point of view of rational types, in Proposition 1, commitment types are simply replaced by types who follow the same plans as their unique rationalizable plan. Hence, under any rationalizable solution concept, the rational types face the same strategic uncertainty in both games leading the same set of possible behavior. We will next establish such strategic equivalence formally.

5. STRATEGIC EQUIVALENCE

In this section, under a broad class of solution concepts, we show that, in Proposition 1, the elaborations $G$ with commitment types and $G'$ without commitment types are strategically equivalent for rational types, in that the set of solutions for each rational type are

\[^5\text{Note also that the technique we use in transforming the model without common-prior to the one with common prior also renders } \pi^c(\tau_1^c) \text{ small, bringing } p^c \text{ near 1 even when } \varepsilon \text{ and } \varepsilon' \text{ are far apart.}\]
identical in games $G$ and $G'$. Therefore, the same set of behavior can be supported by reputational models regardless of whether one allows commitment types. In other words, the same set of behavior is supported whether one allows payoff functions that are inconsistent with the repeated-game structure or imposes this structure throughout. Clearly, our result here applies to any solution concept that is invariant to replacing commitment types with types that have unique rationalizable action plans (in reduced form). In general Bayesian games, this invariance condition is somewhat stronger than elimination of non-rationalizable strategies, because the new game contains some new types, encoding the beliefs of the types with unique rationalizable plans. We first establish our result for a general class of such invariant solution concepts. We also establish the same strategic equivalence for sequential equilibrium; this requires an additional off-path belief restriction commonly imposed in the reputation literature.

The following definitions are standard: A solution concept $\Sigma$ maps every Bayesian game $G$ to a set $\Sigma(G)$ of mixed strategies in game $G$. For any type spaces $T$ and $T'$ with $T \subset T'$ and any strategy profile $\sigma$ on $T'$, $\sigma_T$ denotes the restriction of $\sigma$ to $T$. In the following definitions, we also use the convention that two probability distributions that have common support and agree on this support are identical, ignoring any difference in domains.

**Definition 3.** A solution concept $\Sigma$ is said to be invariant to elimination of non-rationalizable strategies if and only if

$$\Sigma(G) = \Sigma(G')$$

for any two games $G$ and $G'$ with identical type spaces such that (i) if an action plan $s_i$ is available for a type $\tau_i$ in game $G$ then $s_i$ is available for $\tau_i$ in $G'$ and (ii) if $s_i$ is not available for $\tau_i$ in $G$ then $s_i \not\in S^\infty_i [\tau_i|G']$.

**Definition 4.** A solution concept $\Sigma$ is said to be invariant to trivial enrichments of the type spaces if and only if

$$\Sigma(G) = \{\sigma_T|\sigma \in \Sigma(G')\}$$

for any two games $G$ and $G'$ with type spaces $T$ and $T'$ such that (i) $T \subset T'$, (ii) every type in $T$ has identical set of available action plans in games $G$ and $G'$, and (iii) any type in $T$ with multiple action plans has identical interim beliefs in games $G$ and $G'$. 
Note that the transformation in the first definition allows only elimination of non-rationalizable actions and the transformation in the second definition allows only inclusion of new types such that the types who put positive probability to the new types are trivial in that they can play only according to one plan. Proposition 1 implies that under any solution concept that is invariant to the above transformations, elaborations with or without commitment types have the same strategic implications to rational types. Due to its importance, we state this corollary as a proposition:

**Proposition 2.** Let $\Sigma$ be a solution concept that is invariant to elimination of non-rationalizable strategies and to trivial enrichment of the type spaces. Then, for any $\varepsilon' > \varepsilon > 0$ and for any $\varepsilon$-elaboration $G$ with commitment types, there exists an $\varepsilon'$-elaboration $G' = (N, A, (G, T, \pi'))$ without commitment types such that

$$\{\sigma (\tau^*) | \sigma \in \Sigma (G)\} = \{\sigma (\tau^*) | \sigma \in \Sigma (G')\},$$

i.e., the set of solutions for rational types are identical in games $G$ and $G'$.

**Proof.** Note that, in Proposition 1, the elaboration $G'$ can be obtained from $G$ by first introducing new types and then allowing commitment types to play any action plan in the repeated game. The conclusion then follows from the invariance conditions in the above definitions.

We will next establish the same strategic equivalence under sequential equilibrium, which is defined as follows. Given any Bayesian repeated game with a type space $(G, T, \pi)$, a **belief structure** is a list $\mu = \left(\mu_{i,\tau_i,h}\right)_{i \in N, \tau_i \in T_i, h \in H}$ of type specific beliefs $\mu_{i,\tau_i,h} \in \Delta (G \times T_{-i})$ regarding the underlying payoffs and the other player’s types, beliefs that vary with the history of play. An assessment is a pair $(\bar{\sigma}, \mu)$ of strategy profile $\bar{\sigma} : T \to \Sigma$ and a belief structure $\mu$. An assessment $(\bar{\sigma}, \mu)$ is said to be sequentially rational if $\bar{\sigma}_i (\cdot | \tau_i)$ is a sequential best response to $\mu_{i,\tau_i,h}$ and $\bar{\sigma}_{-i}$, i.e., the restriction of $\bar{\sigma}_i (\cdot | \tau_i)$ to the continuation game after every history $h$ remains to be a best response to $\bar{\sigma}_{-i}$ and the beliefs $\mu_{i,\tau_i,h}$ in the continuation game. An assessment $(\bar{\sigma}, \mu)$ is said to be consistent if there exists a sequence $(\bar{\sigma}^n, \mu^n) \to (\bar{\sigma}, \mu)$ such that $\bar{\sigma}^n$ assigns positive probability to each available move at every

---

6 A more general definition of a belief structure would also specify the beliefs regarding past actions, but those beliefs are trivial because of perfect monitoring.
history and \( \mu^n \) is derived from Bayes’ rule and \( \tilde{\sigma}^n \). An assessment \((\tilde{\sigma}, \mu)\) is said to be a \textit{sequential equilibrium} if it is sequentially rational and consistent.

In an \( \varepsilon \)-elaboration without commitment types, sequential equilibria are defined as above. In an \( \varepsilon \)-elaboration with commitment types, the definition of course depends on how one formalizes the commitment types. In particular, the Bayesian game considered in the above definition above implies that Player 2 puts probability 1 on the rational type of Player 1 if the history is not consistent with any commitment type—even when the history is inconsistent with the strategy of the rational type. This is because the commitment types have only one action, so that only the rational types may tremble. This is an additional assumption when the commitment types are represented by payoff perturbations (violating the additive repeated game structure). In general, the possible off-the-path beliefs can vary depending on the way the commitment types are formulated, but the above assumption is usually maintained. We will also keep this additional assumption in our definition for sequential equilibrium without commitment types:

\textbf{Assumption 1.} For every history \( h = (a^0, \ldots, a^{t-1}) \),

\[ \mu_{2, \tau_2^*, h} (g^*, \tau_1^*) = 1 \]

whenever

\[ \left\{ \tau_1 \in T_1 \setminus \{\tau_1^*\} : \prod_{t' \leq t} \sigma_1 \left( a_{t'}^+ | h_{t'}^+, \tau_1 \right) > 0 \right\} = \emptyset, \]

where \( h'' = (a')_{t < t'} \) is the subhistory of history \( h \) at the beginning of \( t \).

In our analysis we will focus on the behavior of the rational types under sequential equilibrium, which is formally defined as follows.

\textbf{Definition 5.} For any elaboration \( G \) (with or without commitment types), we write

\[ SE^* (G) = \{ (\sigma (\cdot | \tau^*) ) | (\sigma, \mu) \text{ is a sequential equilibrium of } G \text{ that satisfies Assumption 1} \} \]

for the set of sequential equilibrium action plans for the rational types in \( G \).

We are now ready to state the strategic equivalence result for sequential equilibrium.
Proposition 3. For any $\varepsilon' > \varepsilon > 0$ and for any $\varepsilon$-elaboration $G$ with commitment types $(C, \pi)$ there exists a $\varepsilon'$-elaboration $G' = (N, A, (G, T', \pi'))$ without commitment types such that

$$SE^* (G) = SE^* (G'),$$

i.e., under Assumption [1], the set of sequential equilibrium action plans for the rational types is same in games $G$ and $G'$.

Proof. We will show that both conditions $\sigma (\cdot | \tau^*) \in SE^* (G)$ and $\sigma (\cdot | \tau^*) \in SE^* (G')$ are characterized the following conditions (SR1) and (SR2). First, $(\sigma, \mu)$ is a sequential equilibrium of $G$ if and only if the following three conditions are satisfied. Consistency condition for $\tau^*_2$ is

(C) $\mu_{2, \tau^*_2, h} (c) = \sigma (\cdot | \tau^*_1) (c)$ $\equiv \begin{cases} \frac{\pi (c)}{\Pr (h | \sigma (\cdot | \tau^*_1)) (1 - \varepsilon) + \sum_{c' \in C^h} \pi (c')} & \text{if } c \in C^h \\ 0 & \text{otherwise} \end{cases}$ $(\forall h, c)$

where $C^h$ is the set of commitment plans $c \in C$ that is consistent with history $h$. Of course, $\mu_h \sigma (\cdot | \tau^*_1) (\tau^*_1) = 1 - \sum_{c \in C} \mu_h \sigma (\cdot | \tau^*_1) (c)$. The consistency condition for player 1 is trivial, as player 2 has only one type. Note that $\mu_h \sigma (\cdot | \tau^*_1)$ is a function of $\sigma (\cdot | \tau^*_1)$, and hence the following sequential rationality conditions are solely on $\sigma (\cdot | \tau^*)$. The sequential rationality conditions are

(SR1): $\sigma (\cdot | \tau^*_1)$ is a sequential best response to $\sigma (\cdot | \tau^*_2)$ under $g^*_1$, and

(SR2): at each history $h$, $\sigma (\cdot | \tau^*_2)$ is conditional best response to the mixed strategy

$$\bar{\sigma} \equiv \mu_h \sigma (\cdot | \tau^*_1) (\tau^*_1) \sigma (\cdot | \tau^*_1) + \sum_{c \in C} \mu_h \sigma (\cdot | \tau^*_1) (c) c$$

under $g^*_2$.

Since all the other types are committed to a single plan, there are no other conditions. This shows that $\sigma (\cdot | \tau^*) \in SE^* (G)$ if and only if (SR1) and (SR2) are satisfied.

To show that $\sigma (\cdot | \tau^*) \in SE^* (G')$ implies the conditions (SR1) and (SR2), consider any sequential equilibrium $(\sigma, \mu')$ of $G'$ that satisfies Assumption [1]. Firstly, since type $\tau^*_1$ puts probability one on $(g^*, \tau^*_2)$, the sequential rationality condition for that type is (SR1). Secondly, since $c$ is the unique rationalizable action plan of $\tau^*_1$ in $G'$ (by Lemma [2]) on all histories
h consistent with c,

\[(5.1) \quad \sigma(c(h)|h, \tau_1^c) = 1 \quad (\forall c \in C^h, \forall h).\]

Hence, by Assumption 1 and consistency,

\[(5.2) \quad \mu_{2,\tau_2^c,h}(\tau_1^c) = \mu_h^{\sigma(\cdot|\tau_1^c)}(c) \quad (\forall h, c),\]

which of course also implies that \(\mu_{2,\tau_2^c,h}(\tau_1^c) = \mu_h^{\sigma(\cdot|\tau_1^c)}(\tau_1^c).\) By (5.1) and (5.2), under the belief of type \(\tau_2^c,\) player 1 plays according to \(\tilde{\sigma}\) above, and the sequential rationality condition for type \(\tau_2^c\) is (SR2).

To show that (SR1) and (SR2) are sufficient for \(\sigma(\cdot|\tau^*) \in SE^*(G'),\) take any \(\sigma(\cdot|\tau^*)\) that satisfies (SR1) and (SR2). We will construct a sequential equilibrium \((\sigma, \mu')\) of \(G'\) that satisfies Assumption 1. Set \(\mu_{1,\tau_1^c}(g^c, \tau_2^c) = 1\) and \(\mu_{2,\tau_2^c,h} = \mu_h^{\sigma(\cdot|\tau_1^c)}\). For each \(c \in C,\) consider a sequential equilibrium \((\sigma^c, \mu^c)\) of the game in which the action plan of type \(\tau_2^c\) is fixed as \(\sigma:\ (\cdot|\tau_2^c)\)—as moves of nature, and the type space is \(T_c\) with the interim beliefs in \(G'.\) Set \(\sigma(\cdot|\tau_i) = \sigma^c(\cdot|\tau_i)\) and \(\mu_{i,\tau_i,h} = \mu_{i,\tau_i,h}^c\) for every \(\tau_i \in T_i^c\) and \(c \in C.\) We now show that \((\sigma, \mu')\) is a sequential equilibrium of \(G'\) and satisfies Assumption 1. Since Lemma 2 applies to the case \(g_2^c = 0,\) in which case \(\sigma(\cdot|\tau_2^c)\) is rationalizable for type \(\tau_2^c,\)

\[(5.3) \quad \sigma^c(c(h)|h, \tau_1^c) = 1 \quad (\forall c \in C^h, \forall h).\]

Hence, \(\mu_{2,\tau_2^c,h}\) is consistent and satisfies Assumption 1. The sequential rationality conditions for rational types are (SR1) and (SR2) by construction and (5.3). The sequential rationality and consistency for types in \(T_c\) immediately follows from the construction and the fact that \((\sigma^c, \mu^c)\) is a sequential equilibrium in the auxiliary game.

The strategic equivalence under sequential equilibrium is somewhat subtle, requiring the lengthy proof above. This is because of the issues relating to the off-the-path beliefs, which play a central role in sequential equilibrium while not being relevant for ICR. If a type \(\tau_1^c\), who plans to follow \(c,\) deviates from \(c,\) then his subsequent behavior may be different from \(c\) as ICR cannot restrict the behavior at the contingencies that are precluded by one’s own strategy. In that case, off the path beliefs of player 2 at the histories that are not consistent with any type could be different. Moreover, consistency may result in unforeseen restrictions on those beliefs as it is applied for types in \(T^c\) and \(\tau_2^c\) simultaneously. Assumption 1 ensures that Player 2 assigns zero probability to \(\tau_1^c\) whenever Player 1 deviates from \(c,\) resulting
in beliefs that are identical to those with commitment types, as we show in the proof. Of course, at the histories that are consistent with commitment types, the rational types in the games $G$ and $G'$ face the same uncertainty regarding all relevant aspects, such as whether the other player is rational and which $c \in C$ he is playing if he is not rational. This leads to the same set of solutions for rational types in both games.

6. Necessity of Commitment under CK of Approximate Payoffs

In the previous sections, while we imposed the constraint that it is always common knowledge that the payoffs are the sum of identical stage-game payoffs, we allowed those payoffs to lie anywhere in the interval $[0, 1]$. In this section, by contrast, we make the stricter requirement that it is common knowledge that payoffs lie within $\varepsilon$ of those in the complete-information game. Under this stricter requirement, we show that commitment types are not dispensable in reputation models. When the stage game is dominance solvable, there is a unique sequential Nash equilibrium outcome, in which the unique rationalizable strategy profile of the stage game is played throughout. Here, $\varepsilon$ is uniform over all type spaces and the number of repetitions. For example, in the repeated prisoners’ dilemma, one cannot have any cooperation without commitment types when it is common knowledge that the payoffs are approximately those in the prisoner’s dilemma.

Define the distance between two stage-game payoff functions via the sup norm:

$$d(g', g) = \max_a |g'(a) - g(a)|$$

**Proposition 4.** Fix a complete information stage game $g^*$ which has unique rationalizable profile $a^*$. Then, there exists $\varepsilon > 0$ such that for any $\varepsilon' > 0$ and any $\bar{t}$, every $\varepsilon'$-elaboration $(N, A, (G, T, \pi))$ without commitment types, satisfying the additional requirement that $d(g, g^*) < \varepsilon$ for all $g \in G$, has a unique sequential equilibrium in which $a^*$ is played by all types at all histories.

**Proof.** The elimination process for the finite stage game $g^*$ is finite. Each time an action is eliminated (again by finiteness) it must be that for some $\delta$ it is never within $\delta$ of being a best reply. Choose $\varepsilon$ so that $2\varepsilon$ is smaller than the minimum of these $\delta$. 
Now suppose there is a sequential equilibrium strategy profile \( s^* \) which contradicts the result. Consider one of the latest histories at which any violation of the profile \( a^* \) occurs, and of the violations at this history, consider an action \( a'_i \) which is eliminated first in the elimination process for \( g^* \), say at stage \( k \). When player \( i \) takes this action, he must believe that (a) the profile \( a^* \) is played at all future dates regardless of his action and (b) no action eliminated at stage \( k - 1 \) or earlier is played at the current history. But then by (b), the fact that \( a'_i \) is eliminated at stage \( k \), and the choice of \( \varepsilon \), his action is suboptimal in the stage game; and by (a) his action cannot affect future play. This contradicts the concept of sequential equilibrium.

For example, in a repeated prisoners’ dilemma game, if it is common knowledge that payoffs are close to the prisoners’ dilemma, then in any sequential equilibrium the players defect throughout the game regardless of the number of repetitions. At some level this is a reflection of general continuity properties of Bayesian Nash equilibrium payoffs with respect to the perturbations of payoffs. Indeed, it is well known that, for any given \( \bar{\varepsilon} \), as \( \varepsilon \to 0 \), the Bayesian equilibrium payoffs in \( \varepsilon \)-elaborations of repeated prisoners’ dilemma with commitment types approach the payoffs from defection throughout the game. This is in line with the continuity results for Nash equilibrium payoffs with respect to the prior distributions. Hence, for a given \( \bar{\varepsilon} \), our result here differs from the existing continuity results only in terms of the perturbations it considers, making the stage payoffs approach to the original game instead of making the probability of “crazy” types with unrelated payoffs to go to zero. Our result has a major strength however: \( \varepsilon \) is uniform with respect to the number of repetition. In contrast, for any \( \varepsilon \) probability of a tit-for-tat type, cooperation prevails whenever the number of repetitions are sufficiently large, as famously established by Gang of Four.

For arbitrary stage games, one can further show that commitment types cannot be constructed. In fact, given any move which is strictly rationalizable in the stage game, the plan which repeats this move at all histories is always rationalizable for any sufficiently small \( \varepsilon \). Hence there cannot be commitment types which find general plans uniquely rationalizable.
7. Conclusion

Appendix A. Omitted Proofs

A.1. Preliminary Definitions. In the appendix, we will also consider type spaces without a common prior. Such a type space is a list \((G, T, \pi(\cdot|\cdot))\) where \(\pi(\cdot|\tau_i) \in \Delta (G \times T)\) is the probability distribution of \(\tau_i\). Here, there need not be a single \(\pi \in \Delta (G \times T)\) that leads to these interim beliefs by Bayes' rule. Fix any \(G = (N, A_i,G, T, \pi(\cdot|\cdot))\). For each \(i \in N\) and for each belief \(\beta \in \Delta (G \times S_{-i})\), we write \(BR_i(\beta)\) for the set of actions \(s_i \in S_i\) that maximize the expected value of \(u_i(z(s_i, s_{-i})|g)\) under the probability distribution \(\beta\).

Interim correlated rationalizability (ICR) is computed by the following elimination procedure. For each \(i\) and \(\tau_i\), set \(S^0_i[\tau_i|G] = S_i\), and define sets \(S^k_i[\tau_i|G]\) for \(k > 0\) iteratively, by letting \(s_i \in S^k_i[\tau_i|G]\) if and only if \(s_i \in BR_i \left(\text{marg}_{G \times S_{-i}} \beta\right)\) for some \(\beta \in \Delta (G \times T_{-i} \times S_{-i})\) such that \(\text{marg}_{G \times T} \beta = \pi(\cdot|\tau_i)\) and \(\beta(s_{-i} \in S^{k-1}_{-i}[\tau_{-i}|G]) = 1\). That is, \(s_i\) is a best response to a belief of \(\tau_i\) that puts positive probability only to the actions that survive the elimination in round \(k - 1\). We write \(S^k[\tau|G] = S^k_i[\tau_i|G] \times S^k_S[\tau_2].\) Then,

\[
S^\infty_i[\tau_i|G] = \bigcap_{k=0}^{\infty} S^k_i[\tau_i|G].
\]

The following class of action plans will play an important role in our construction:

**Definition 6.** A plan \(s_i\) is said to be sure-thing compliant if and only if there is no partial history \(h\) and move \(a_i \in A_i\) such that \(s_i(h, (s_i(h), a_{-i})) = a_i\) for every \(a_{-i}\) but \(s_i(h) \neq a_i\).

In other words, a plan is sure-thing compliant if whenever the player plays \(a_i\) in all possible continuations next period, he also plays \(a_i\) this period. This is of course equivalent to the sure-thing principle of Savage if the player has the same preferences over his moves in both periods.

We will finally introduce couple notation regarding equivalent action plans. Recall that \(z(s)\) denotes the outcome of a profile \(s\) of action plans. In line with our notation for histories, we will write \(z(s)^t\) for the truncation of \(z(s)\) at the beginning date \(t\); i.e., if \(z(s) = (a^0, a^1, \ldots , a^t)\), then \(z(s)^t = (a^0, a^1, \ldots , a^{t-1})\). Recall also that action plans \(s_i\) and \(s'_i\) are equivalent if \(z(s_i, s_{-i}) = z(s'_i, s_{-i})\) for all action plans \(s_{-i} \in S_{-i}\), i.e., they lead to the same outcome no matter what strategy the other player plays. Note that \(s_i\) and \(s'_i\) are equivalent iff \(s_i(h^i) = s'_i(h^i)\) for every history \(h^i\) in which \(i\) played according to \(s_i\) throughout; they may differ only in their prescriptions for histories.
that they preclude. Hence, in reduced form, action plans can be represented as mappings that maps the history of other players’ play into own stage game actions. Similarly, action plans $s_i$ and $s_i'$ are said to be $t$-equivalent if $z(s_i, s_{-i})^t = z(s_i', s_{-i})^t$ for all action plans $s_{-i} \in S_{-i}$, they lead to the same history up to date $t$ no matter what strategy the other player plays. Equivalence and $t+1$-equivalence are equivalent. Given any two sets $X, Y$ of action plans, we write $X \simeq Y$ if for every $x \in X$ there exists $y \in Y$ that is equivalent to $x$, and for every $y \in Y$ there exists $x \in X$ that is equivalent to $x$. We define the relation $X \simeq^t Y$ similarly for $t$-equivalence.

A.2. Proof of Lemma 1. Our proof has three main steps. First, we will prove it for sure-thing compliant action plans without requiring the type space to have common prior or the full support property, the property 1 in the statement of the lemma. We then extend this result to all action plans, without requiring the properties on type space once again. Finally, we convert the latter type space to a type space with common prior and full support assumptions without altering the rationalizable actions, proving the lemma. The first step is the following lemma; Weinstein and Yildiz (2012) proves this lemma for infinite-horizon games, and the proof carries over to the finitely repeated games without any major modification. (We include the proof for now for completeness.)

**Lemma 3** (Weinstein Yildiz, 2012). For any sure-thing compliant action plan $s_i$ and any $L \in T$, there exists a game $\tilde{G} = (N, A, (\tilde{G}, \tilde{T}, \tilde{\pi} (\cdot | \cdot)))$ with a type $\tau^s_{i,L}$ such that $S^\infty \left[ \tau^s_{i,L} | \tilde{G} \right] \simeq^L \{ s_i \}$. (The type space does not necessarily have a common prior.)

**Proof.** We will induct on $L$. When $L = 1$, it suffices to consider a type $\tau^{s_i,L}$ who is certain that in the stage game, $s_i(\emptyset)$ yields payoff 1 while all other actions yield payoff 0. Now fix $L, s_i$ and assume the result is true for all players and for $L - 1$. In outline: the type we construct will have payoffs which are completely insensitive to the actions of the other players, but will find those actions informative about his own payoffs. He also will believe that if he ever deviates from $s_i$, the other players’ subsequent actions are uninformative — this ensures that he always chooses the myopically best action.

Formally: Let $\hat{H}$ be the set of histories of length $L - 1$ in which player $i$ always follows the plan $s_i$, so that $|\hat{H}| = |A_{-i}|^{L-1}$, where $A_{-i}$ is the set of profiles of static moves for the other players. For each history $h \in \hat{H}$, we construct a pair $(\tau^h_{i,L}, g^h)$, and our constructed type $\tau^{s_i,L}$ assigns equal weight to each of $|A_{-i}|^{L-1}$ such pairs. Each type $\tau^h_{i,L}$ is constructed by applying the inductive hypothesis to a plan $s^h_{-i}$ which plays according to history $h$ so long as $i$ follows $s_i$, and simply repeats the previous move forever if player $i$ deviates. Such plans are sure-thing compliant for the player $-i$ because at every history, the current action is repeated on at least one branch.
To define the payoff functions \( \theta^h \) for all \( h \in \hat{H} \), we will need to define an auxiliary function \( f : \hat{H} \times A_i \rightarrow \mathbb{R} \), where \( \hat{H} \) is the set of prefixes of histories in \( \hat{H} \). The motive behind the construction is that \( f(h, \cdot) \) represents \( i \)'s expected value of his stage-game payoffs conditional on reaching the history \( h \). The function \( f \) is defined iteratively on histories of increasing length. Specifically, define \( f \) as follows: Fix \( \varepsilon > 0 \). Let \( f(\emptyset, s_i(\emptyset)) = 1 \) and \( f(\emptyset, a) = 0 \) for all \( a \neq s_i(\emptyset) \), where \( \emptyset \) is the empty history. Next, assume \( f(h, \cdot) \) has been defined and proceed for the relevant one-step continuations of \( h \) as follows:

Case 1: If \( s_i(h, (s_i(h), a_{-i})) = s_i(h) \) for all \( a_{-i} \), then let \( f((h, a), \cdot) = f(h, \cdot) \) for every \( a \).

Case 2: Otherwise, by sure-thing compliance, at least two different actions are prescribed for continuations \( (h, (s_i(h), a_{-i})) \) as we vary \( a_{-i} \). For each action \( a_i \in A_i \), let \( S_{a_i} = \{ a_{-i} : s_i(h, (s_i(h), a_{-i})) = a_i \} \) be the set of continuations where \( a_i \) is prescribed. Then let

\[
 f((h, (s_i(h), a_{-i})), a_i) = \begin{cases} 
 f(h, s_i(h)) + \varepsilon & \text{if } a_{-i} \in S_{a_i} \\
 \frac{|A_{-i}| f((h, a_{-i} \setminus s_i(h))) - |S_{a_i}| f((h, s_i(h)) + \varepsilon)}{|A_{-i}| - |S_{a_i}|} & \text{if } a_{-i} \notin S_{a_i}
\end{cases}
\]

where the last denominator is non-zero by the observation that at least two different actions are prescribed.

These payoffs were chosen to satisfy the constraints

\[
(A.1) \quad f(h, a_i) = \frac{1}{|A_{-i}|} \sum_{a_{-i}} f(h, (s_i(h), a_{-i})), a_i
\]

\[
(A.2) \quad f(h, s_i(h)) \geq f(h, a_i) + \varepsilon \quad (\forall h, a_i \neq s_i(h)).
\]

as can be verified algebraically.

For each history \( h \in \hat{H} \), define the stage-game payoff function \( g^h : A \rightarrow [0, 1]^n \) by setting \( g^h_i(a) = f(h, a_i) \) and \( g^h_j(a) = 0 \) at each \( a \) and \( j \neq i \). Define \( \tau^{s_i,L} \) as mentioned above, by assigning equal weight to each pair \((\tau^{h_{-i}}, \theta^h)\).

We claim that under rationalizable play, from the perspective of type \( \tau^{s_i,L} \), when he has followed \( s_i \) and reaches history \( h \in \hat{H} \), \( f(h, \cdot) \) is his expected value of the stage-game payoff \( g_i \). We show this by induction on the length of histories, backwards. When a history \( h \in \hat{H} \) is reached, player \( i \) knows (assuming rationalizable play) the opposing types must be \( \tau^{h_{-i}} \) and thus the stage-game payoff function must be \( g^h \), which is the desired result for this case. Suppose the claim is true for all histories in \( \hat{H} \) of length \( M \). Note that type \( \tau^{s_i,L} \) puts equal weight on all sequences of play for his opponent. Therefore, for a history \( h \in \hat{H} \) of length \( M - 1 \), the expected payoffs are given by the right-hand-side of (A.1) which proves the claim.
Note also that if he follows \( s_i \) through period \( L \), player \( i \) always learns his true payoff. Let \( \tilde{s}_i \) be the plan which follows \( s_i \) through period \( L \), then plays the known optimal action from period \( L + 1 \) onward. We claim that \( \tilde{s}_i \) strictly outperforms any plan which deviates by period \( L \). The intuitive argument is as follows. Because type \( \tau^{s_i,L} \) has stage-game payoffs which are insensitive to the other players’ moves, he only has two possible incentives at each stage: the myopic goal of maximizing his average stage-game payoffs at the current stage, and the desire to receive further information about his payoffs. The former goal is strictly satisfied by the move prescribed by \( \tilde{s}_i \), and the latter is at least weakly satisfied by this move, since after a deviation he receives no further information.

Formally, we must show that for any fixed plan \( s'_i \) not \( L \)-equivalent to \( s_i \) and any rationalizable belief of \( \tau^{s_i,L} \), the plan \( \tilde{s}_i \) gives a better expected payoff. Given a rationalizable belief on opponents’ actions, player \( i \) has a uniform belief on the other players’ actions as long as he follows \( s_i \). Let \( \hat{h} \) be a random variable equal to the shortest realized history at which \( s'_i \) differs from \( s_i \) before period \( L \), or \( \infty \) if they do not differ by period \( L \). Note that the uniform belief on others’ actions implies that \( \hat{h} \neq \infty \) with positive probability. We show that conditional on any non-infinite value of \( \hat{h} \), \( \tilde{s}_i \) strictly outperforms \( s'_i \) on average. In fact this is weakly true stage-by-stage, and strictly true at the first deviation, because:

At stages \( 1, \ldots, |\hat{h}| \): The plans are identical.

At stage \( |\hat{h}| + 1 \): The average payoff \( f(\hat{h}, a_i) \) is strictly optimized by \( \tilde{s}_i(\hat{h}) \).

At stages \( |\hat{h}| + 2, \ldots, L \): Along the path observed by a player following \( s'_i \), the other players are known to repeat their stage-\( |\hat{h}| + 1 \) move at stages \( |\hat{h}| + 2, \ldots, L \). So at these stages, the plan \( s'_i \) cannot do better than to optimize with respect to the history truncated at length \( |\hat{h}| + 1 \). The plan \( \tilde{s}_i \) optimizes the expected stage-game payoffs with respect to a longer history, under which opposing moves are identical through stage \( |\hat{h}| + 1 \). Since he is therefore solving a less-constrained optimization problem, he must perform better than \( s'_i \) at each stage \( |\hat{h}| + 2, \ldots, L \).

At stages \( L + 1, \ldots \): Under plan \( \tilde{s}_i \), player \( i \) now has complete information about his payoff and optimizes perfectly, so \( s'_i \) cannot do better.

If \( \hat{h} = \infty \), again \( \tilde{s}_i \) cannot be outperformed because he optimizes based on complete information after \( L \), and \( \tilde{s}_i \) and \( s'_i \) prescribe the same behavior before \( L \).

Finally, since there are only finitely many histories and types in the construction, all payoffs are bounded and so can be normalized to lie in \([0, 1]\). □

The next lemma builds on this result to generalize to all action plans.
Lemma 4. For any action plan $s_i$, there exists a game $\hat{G} = (N, A, (\hat{G}, \hat{T}, \hat{\pi} (\cdot | \cdot)))$ with a type $\tau_i^{s_i}$ such that $S_i^{\infty} [\tau_i^{s_i} | \hat{G}] \simeq \{s_i\}$. (The type space does not necessarily have a common prior.)

Proof. Fix some $a^*_i \in A_{-i}$, and define a function $v_{-i} : A_{-i} \to [0, 1]$ by

$$v_{-i} (a_{-i}) = \begin{cases} 1 & \text{if } a_{-i} = a^*_i, \\ 0 & \text{otherwise.} \end{cases}$$

For every $\hat{a}_i \in A_i$, define also a function $v_{i}^{\hat{a}_i} : A_i \to [0, 1]$ by

$$v_{i}^{\hat{a}_i} (a_i) = \begin{cases} 1 & \text{if } a_i = \hat{a}_i, \\ 0 & \text{otherwise.} \end{cases}$$

Finally, consider the class of stage-game payoʃ functions $g_i^{\hat{a}_i} : A \to [0, 1]$ for player $i$ where

$$g_i^{\hat{a}_i} (a_i, a_{-i}) = \lambda v_i^{\hat{a}_i} (a_i) + (1 - \lambda) v_{-i} (a_{-i})$$

for some $\lambda \in (0, 1/(2\tilde{t} + 1))$. Let also $\hat{S}_{-i}$ be the set of action proﬁles $s_{-i}$ for which there exists a function $\rho : H \times A_{-i} \to A_i$ such that

1. for any $t < \tilde{t}$, any history $h^t$ and any $(a_i, a_{-i}) \in A_i$,

$$s_{-i} (h^t, a_{-i}) = a^*_i \iff a_i = \rho (h^t, a_{-i});$$

2. $\rho (h^{\tilde{t}-1}, a_{-i}) \equiv s_i (h^{\tilde{t}-1})$ for all those $h^{\tilde{t}-1}$ such that player $i$ has played according to $s_i$ throughout.

At any history $(h, (a_i, a_{-i}))$, player $-i$ rewards $i$ by playing $a^*_i$ if $a_i = \rho (h, a_{-i})$. The only restriction on which move is rewarded occurs at date $\tilde{t} - 1$ and in the contingency that $i$ has followed $s_i$ up to $\tilde{t} - 1$: he will be rewarded at $\tilde{t}$ if he continues to follow $s_i$ at $\tilde{t} - 1$. The set $\hat{S}_{-i}$ is symmetric in all other ways. Note also that player $j \neq i$ reacts differently to the rewarded move of player $i$. Hence, the actions in $\hat{S}_{-i}$ are all sure-thing compliant. Thus, by Lemma 3, for each $s_{-i} \in \hat{S}_{-i}$, there exists a game $G^{s_{-i}} = (N, A, (G^{s_{-i}}, T^{s_{-i}}, \pi^{s_{-i}} (\cdot | \cdot)))$ with a type $\tau_i^{s_{-i}}$ such that $S_{-i}^{\infty} [\tau_i^{s_{-i}} | G^{s_{-i}}] \simeq \{s_{-i}\}$. 


Define the game \( \tilde{G} = \left( N, A, (\tilde{g}, \tilde{T}, \tilde{\pi}(\cdot|\cdot)) \right) \) by
\[
\tilde{g} = \left\{ g_{i}^{a_{i}} | a_{i} \in A_{i} \right\} \cup \bigcup_{s_{-i} \in \hat{S}_{-i}} G_{s_{-i}}^{s_{-i}};
\]
\[
\tilde{T}_{i} = \left\{ \tau_{i}^{s_{i}} \right\} \cup \bigcup_{s_{-i} \in \hat{S}_{-i}} T_{i}^{s_{-i}}; \quad \tilde{T}_{-i} = \bigcup_{s_{-i} \in \hat{S}_{-i}} T_{-i}^{s_{-i}};
\]
\[
\tilde{\pi}(\cdot|\tau_{j}) = \pi^{s_{-i}}(\cdot|\tau_{j}) \quad \left( \forall \tau_{j} \in T_{j}^{s_{-i}}, j \in N, s_{-i} \in \hat{S}_{-i} \right);
\]
(A.3) \[
\tilde{\pi}\left(g^{s_{i}}\left(z(s_{i}, s_{-i})^{t}\right), \tau_{-i}^{s_{-i}} \mid \tau_{i}^{s_{i}}\right) = 1/|\hat{S}_{-i}| \quad \left( \forall s_{-i} \in \hat{S}_{-i} \right).
\]

The main attraction of the definition is the belief of the newly introduced type \( \tau_{i}^{s_{i}} \) in [A.3]. He assigns equal probabilities on types \( \tau_{-i}^{s_{-i}} \) but also believes that there is a perfect correlation between the types \( \tau_{-i}^{s_{-i}} \) and the way his own action affects his payoff. If he follows \( s_{i} \) throughout \( h^{\tilde{t}} \) and observes the moves of the other player, he learns what action \( \tilde{a}_{i} \) is best for him, which happens to be the action that he would have played at that history according to \( s_{i} \). Note that each of the types other than \( \tau_{i}^{s_{i}} \) has a unique rationalizable action in reduced form. Hence, \( \tau_{i}^{s_{i}} \) has a unique rationalizable belief regarding the payoff functions and the outcomes.

In the rest of the proof, we will show that \( s_{i} \) is uniquely rationalizable for \( \tau_{i}^{s_{i}} \) in reduced form, i.e., \( S_{i}^{\infty} \left[ \tau_{i}^{s_{i}} | \tilde{G} \right] \simeq \{ s_{i} \} \). We now introduce couple of notations. For any history \( h^{t} \), write \( P_{t}^{*}(h^{t}) \) for the probability that \( a_{-i}^{*} \) is played at date \( t \) conditional on \( h \) according to the rationalizable belief of \( \tau_{i}^{s_{i}} \). As noted above, by symmetry,

\[
P_{t}^{*}(h^{t}) = \begin{cases} 
1 & \text{if } t = \tilde{t} \text{ and } i \text{ follows } s_{i} \text{ throughout } h^{t}; \\
0 & \text{if } t = \tilde{t} \text{ and } i \text{ follows } s_{i} \text{ up to } \tilde{t} - 1 \text{ but deviates at } \tilde{t} - 1 \text{ in } h^{t}; \\
1/|A_{-i}| & \text{otherwise.}
\end{cases}
\]

Write \( U_{i}(s_{i}^{t}|h) \) for the expected payoff of \( i \) from playing \( s_{i}^{t} \) under the rationalizable belief of type \( \tau_{i}^{s_{i}} \) conditional on history \( h \). Write also \( U_{i}(h^{t}) \) for the realized expected payoff of \( \tau_{i}^{s_{i}} \) up to date \( t \) at history \( h^{t} \).

We now show that \( U_{i}(s_{i}|h^{t}) > U_{i}(s_{i}^{t}|h^{t}) \) for every history \( h^{t} \) and action plan \( s_{i}^{t} \) such that \( i \) follows \( s_{i} \) throughout \( h^{t} \) and \( s_{i}^{t}(h^{t}) \neq s_{i}(h^{t}) \). So long as he follows \( s_{i} \), every such history \( h^{t} \) is reached with positive probability under the rationalizable belief of \( \tau_{i}^{s_{i}} \). This therefore shows that the expected payoff from \( s_{i} \) is strictly higher than any \( s_{i}^{t} \) that is not equivalent to \( s_{i} \). Therefore, \( S_{i}^{\infty} \left[ \tau_{i}^{s_{i}} | \tilde{G} \right] \simeq \{ s_{i} \} \).

First consider the case \( t = \tilde{t} \). Conditional on \( h^{t} \), \( \tau_{i}^{s_{i}} \) assigns probability 1 on \( g^{s_{i}}(h^{t}) \). If he follows \( s_{i} \) and plays \( s_{i}(h^{t}) \) at \( h^{t} \), then his own action contributes \( \lambda \) to his payoff, while his own action
contributes zero to his payoff otherwise. Moreover, since he has followed \( s_i \) throughout \( h^t \), he will be rewarded for sure by the other player at \( \bar{t} \), contributing \( 1 - \lambda \) to his payoff regardless of his own move at \( h^t \). Hence,

\[
U_i (s_i|h^t) = U_i (h^t) + 1,
\]

and

\[
U_i (s'_i|h^t) = U_i (h^t) + 1 - \lambda,
\]

yielding

\[
U_i (s_i|h^t) - U_i (s'_i|h^t) = \lambda > 0.
\]

Now consider the case \( t < \bar{t} \). His payoff from following \( s_i \) is

\[
U_i (s_i|h^t) = U_i (h^t) + \lambda \sum_{t' = t}^\bar{t} E \left[ v_i^\hat{s}_i \left( s_i \left( h^{t'} \right) \right) \right] |h^t, s_i] + (1 - \lambda) \sum_{t' = t}^\bar{t} E \left[ P_{t'}^* \left( h^{t'} \right) \right] |h^t, s_i]
\]

\[
\geq U_i (h^t) + (1 - \lambda) \frac{\bar{t} - t}{|A_{-i}|} + 1.
\]

To see the lower bound, note that, so long as he follows \( s_i \), he gets 1 at date \( \bar{t} \) (as in the previous case) and at least \((1 - \lambda)/|A_{-i}| \) at each \( t' < \bar{t} \). (At any \( t' < \bar{t}, v_i^\hat{s}_i \geq 0 \) and \( P_{t'}^* \geq 1/|A_{-i}| \) when he follows \( s_i \).) On the other hand, his payoff from following \( s'_i \) is

\[
U_i (s'_i|h^t) = U_i (h^t) + \lambda \sum_{t' = t}^\bar{t} E \left[ v_i^\hat{s}_i \left( s'_i \left( h^{t'} \right) \right) \right] |h^t, s'_i] + (1 - \lambda) \sum_{t' = t}^\bar{t} E \left[ P_{t'}^* \left( h^{t'} \right) \right] |h^t, s'_i]
\]

\[
\leq U_i (h^t) + \lambda (\bar{t} - t + 1) + (1 - \lambda) \frac{\bar{t} - t + 1}{|A_{-i}|}.
\]

The upper bound comes from the fact that \( v_i^\hat{s}_i \leq 1 \) throughout and \( P_{t'}^* \leq 1/|A_{-i}| \) after a deviation from \( s_i \) (by \( A.4 \)). Combining the two inequalities, we obtain

\[
U_i (s_i|h^t) - U_i (s'_i|h^t) \geq (1 - \lambda) \left( 1 - \frac{1}{|A_{-i}|} \right) - \lambda (\bar{t} - t) > 0,
\]

where the strict inequality follows from \( \lambda < 1/(2\bar{t} + 1) \) and \( |A_{-i}| \geq 2 \).

\[\square\]

**Proof of Lemma 7.** By Lemma 4, there exists a game \( \hat{G} = \left( N, A, \left( \hat{G}, \hat{T}, \hat{\pi} (\cdot | \cdot) \right) \right) \) with a type \( \hat{r}_i \) such that \( S_i^\infty \left( \hat{r}_i | \hat{G} \right) \sim \{s_i\} \). This falls short of the conditions of Lemma 1 in that \( \left( \hat{G}, \hat{T}, \hat{\pi} (\cdot | \cdot) \right) \) does not necessarily admit a common prior and the prior could not have a full support (Condition 1) even if existed. Here, we remedy this problem by converting \( \hat{G} \) to a common prior
game $G^s = (N, A, (G^s, T^s, \pi^s))$ with the desired properties. First, for every $\lambda \in (0, 1)$, define $G^\lambda = \left(N, A, (\tilde{G}, \tilde{T}, \pi^\lambda (\cdot | \cdot)) \right)$ by setting

$$\pi^\lambda (g, \tau_{-j} | \tau_j) = \frac{\lambda}{|\tilde{G} \times \tilde{\tau}_{-j}|} + (1 - \lambda) \tilde{\pi} (g, \tau_{-j} | \tau_j)$$

at each $(g, \tau_j, \tau_{-j}) \in \tilde{G} \times \tilde{T}$. Now, as $\lambda \to 0$, $\pi^\lambda (g, \tau_{-j} | \tau_j) \to \tilde{\pi} (g, \tau_{-j} | \tau_j)$ everywhere. Together with a continuity result for belief hierarchies by Mertens and Zamir (1985), this implies that the belief hierarchy of type $\tilde{\tau}_i$ in game $G^\lambda$ converges to the belief hierarchy of $\tilde{\tau}_i$ in game $\tilde{G}$. Thus, by upperhemicontinuity of ICR (Dekel, Fudenberg, and Morris, 2006), there exists $\lambda > 0$ such that

$$S^\infty_i [\tilde{\tau}_i | G^\lambda] \subseteq S^\infty_i [\tilde{\tau}_i | \tilde{G}] \simeq \{ s_i \}.$$  

Since $S^\infty_i [\tilde{\tau}_i | G^\lambda]$ is non-empty, this implies that

(A.5) $$S^\infty_i [\tilde{\tau}_i | G^\lambda] \simeq \{ s_i \}.$$  

Moreover, since $\tilde{G} \times \tilde{T} \times S$ is finite, there exists some finite $k$ such that

(A.6) $$S^\infty_i [\tilde{\tau}_i | G^\lambda] = S^k_i [\tilde{\tau}_i | G^\lambda].$$  

Now, since $\pi^\lambda (g, \tau_{-j} | \tau_j) > 0$ everywhere, by the main result of Lipman (2003), there exists a common-prior game $\tilde{G}^s = (N, A, (\tilde{G}^s, \tilde{T}^s, \pi^s))$ such that the common prior $\pi^s_i$ is positive everywhere and there exists a type $\tau^s_i \in \tilde{T}^s_i$ whose first $k$ orders of beliefs are identical to that of type $\tilde{\tau}_i$ in game $G^\lambda$. But by Dekel, Fudenberg, and Morris (2007), $S^k$ is a function of the first $k$ orders of beliefs, yielding

(A.7) $$S^k_i [\tau^s_i | G^s] = S^k_i [\tilde{\tau}_i | G^\lambda].$$  

Combining (A.5), (A.6) and (A.7), we obtain

$$S^\infty_i [\tau^s_i | G^s] \subseteq S^k_i [\tau^s_i | G^s] = S^k_i [\tilde{\tau}_i | G^\lambda] \simeq \{ s_i \}. $$

Since $S^\infty_i [\tau^s_i | G^s] \neq \emptyset$, this further implies that

$$S^\infty_i [\tau^s_i | G^s] \simeq \{ s_i \},$$

as desired. □
A.3. Proof of Lemma 2. Using induction on \( k \), we will show that \( S^k_{-i} [\tau_i|G] = S^k_{-i} [\tau_i|G^c] \) for every \( k, \tau_i \in T^i \), and \( i \in N \). This is true for \( k = 0 \) by definition. Towards an induction, assume that

\[
S^{k-1}_{-i} [\tau_{-i}|G] = S^{k-1}_{-i} [\tau_{-i}|G^c] \quad (\forall \tau_{-i} \in T^{i}_{-i}) .
\]

Take any \( \tau_i \in T^i \) and write \( B(\tau_i|G) \) for the set of all beliefs \( \beta \) of type \( \tau_i \) after round \( k-1 \) in game \( G \) for any \( G \in \{G', G^c\} \), where \( \text{marg}_{G \times T_{-i}} \beta = \pi (|\tau_i) \) and \( \beta \left( s_{-i} \in S^{k-1}_{-i} [\tau_{-i}|G] \right) = 1 \). First consider the case \( \tau_i \neq \tau_i^c \). In that case, by definition, \( \pi' (|\tau_i) = \pi (|\tau_i) \). Together with the inductive hypothesis \( \text{(A.8)} \), this implies that \( B(\tau_i|G') = B(\tau_i|G^c) \). Therefore, \( \tau_i \in S^k_{-i} [\tau_i|G^c] \) if and only if \( \tau_i \in S^k_{-i} [\tau_i|G^c] \), and this is the case if and only if \( s_i \in S_k^{i} [\tau_i|G^c] \), showing that \( S^k_{-i} [\tau_i|G^c] = S^k_{-i} [\tau_i|G^c] \).

Now consider the case \( \tau_i = \tau_i^c \). Then,

\[
\pi' (|\tau_i) = p^c \delta_{(\mathbf{0}, g^*_2)} (\tau_2^c) + (1 - p^c) \pi (|\tau_i)
\]

where the probability \( p^c \in (0, 1) \) is defined in \( \text{(4.1)} \), and \( \delta_x \) is the Dirac’s measure on \( x \), putting probability 1 on \( \{x\} \). Hence, by the inductive hypothesis \( \text{(A.8)} \), \( \beta \in B(\tau_i|G^c) \) if and only if

\[
\beta (|T_2^c) \in B(\tau_i|G^c).
\]

Now, take any \( s_i \in S^k_{-i} [\tau_i|G^c] \). Then, \( s_i \in BR_i \left( \text{marg}_{G' \times S_{-i} \beta} \right) \) for some \( \beta \in B(\tau_i|G^c) \). By \( \text{(A.10)} \), for any \( s'_i \),

\[
p^c \cdot 0 + (1 - p^c) E_{\beta (|T_2^c)} [u_i (s_i, s_{-i}|g)] = E_{\beta} [u_i (s_i, s_{-i}|g)] \geq E_{\beta} [u_i (s'_i, s_{-i}|g)] = p^c \cdot 0 + (1 - p^c) E_{\beta (|T_2^c)} [u_i (s'_i, s_{-i}|g)]
\]

where \( \beta (|T_2^c) \in B(\tau_i|G^c) \). (Here, the inequality follows from \( s_i \) being a best response, and the equalities follow from \( \text{(A.10)} \).) Since \( p^c < 1 \), this further implies that

\[
E_{\beta (|T_2^c)} [u_i (s_i, s_{-i}|g)] \geq E_{\beta (|T_2^c)} [u_i (s'_i, s_{-i}|g)] ,
\]

showing that \( s_i \in BR_i \left( \text{marg}_{G' \times S_{-i} \beta} (|T_2^c) \right) \). Therefore, \( s_i \in S^k_{-i} [\tau_i|G^c] \).

Conversely, take any \( s_i \in S^k_{-i} [\tau_i|G^c] \). By definition, \( s_i \in BR_i \left( \text{marg}_{G' \times S_{-i} \beta} (|T_2^c) \right) \) for some \( \beta (|T_2^c) \in B(\tau_i|G^c) \). Pick any \( \beta (|T_2^c) \in \Delta \left( \{((\mathbf{0}, g^*_2), \tau_2^c)\} \times S^{k-1}_{-i} [\tau_2^c|G^c]\right) \), and define \( \beta \in
REPUTATION WITHOUT COMMITMENT

$B(\tau_i|G^r)$ by (A.10). Now, for any $s'_i$,

$E_\beta[u_i(s_i, s_{-i}|g)] = p^c \cdot 0 + (1 - p^c) E_\beta(\tau^*_g)[u_i(s_i, s_{-i}|g)]$

$\geq p^c \cdot 0 + (1 - p^c) E_\beta(\tau^*_{G^r})[u_i(s'_i, s_{-i}|g)]$

$= E_\beta[u_i(s'_i, s_{-i}|g)],$

where the inequality follows from $s_i$ being a best response, and the equalities follow from (A.10). That is, $s_i \in BR_i \left( \text{ marg}_{G^r \times S_{-i}} \beta \right)$. Therefore, $s_i \in S^k_i [\tau_r|G^r]$.

REFERENCES


WEINSTEIN: NORTHWESTERN UNIVERSITY; YILDIZ: MIT