Forward Induction in a Backward Inductive Manner*

Martin Meier† and Andrés Perea‡

This version: December 2023

Abstract

We propose a new rationalizability concept for dynamic games with imperfect information, *forward and backward rationalizability,* that combines elements from forward and backward induction reasoning. It proceeds by applying the forward induction concept of *strong rationalizability* (also known as *extensive-form rationalizability*) in a backward inductive fashion: It first applies strong rationalizability from the last period onwards, subsequently from the penultimate period onwards, keeping the restrictions from the last period, and so on, until we reach the beginning of the game. We argue that, compared to strong rationalizability, the new concept provides a more compelling theory for how players react to surprises. We show that the new concept always exists, and is characterized epistemically by (a) first imposing common strong belief in rationality from the last period onwards, then (b) imposing common strong belief in rationality from the penultimate period onwards, keeping the restrictions imposed by (a), and so on. It turns out that in terms of outcomes, the concept is equivalent to the pure forward induction concept of strong rationalizability, but both concepts may differ in terms of strategies. In terms of strategies, the new concept provides a refinement of the pure backward induction reasoning as embodied by *backward dominance* and *backwards rationalizability.* In fact, the new concept can be viewed as a backward looking strengthening of the forward looking concept of backwards rationalizability. Combining our results yields that every strongly rationalizable outcome is also backwards rationalizable. Finally, it is shown that the concept of forward and backward rationalizability satisfies the principle of *supergame monotonicity*: If a player learns that the game was actually preceded by some moves he was initially unaware of, then this new information will only refine, but never completely overthrow, his reasoning. Strong rationalizability violates this principle.

*We thank Pierpaolo Battigalli, Adam Brandenburger, Emiliano Catonini, Satoshi Fukuda, Aviad Heifetz, David Pearce and Klaus Ritzberger for valuable comments. In addition, we are grateful to some anonymous referees at LOFT 2022 for their useful feedback. We also thank the audiences at the workshop "10 Years of Epicenter", the conference LOFT 2022, the seminar at Royal Holloway University of London, the seminar at NYU Abu Dhabi, and the SAET Conference 2023 for their comments and suggestions.

†*E-mail:* M.Meier@bath.ac.uk *Affiliation:* University of Bath, Department of Economics, United Kingdom, and IHS Vienna, Austria.

‡*E-mail:* a.perea@maastrichtuniversity.nl *Affiliation:* Maastricht University, EpiCenter and Department of Quantitative Economics, The Netherlands.

JEL Classification: C72

Keywords: Forward induction, backward induction, extensive-form rationalizability, backwards rationalizability, backward dominance, supergame monotonicity, dynamic games

1 Introduction

The main feature that distinguishes dynamic games from static games is the issue of *belief revision.* That is, how does a player revise his belief upon observing a surprising move by an opponent? Backward induction and forward induction reasoning provide two different answers to this question.

The older idea is backward induction, which dates at least back to von Neumann and Morgenstern (1953)[1], and has also been incorporated in concepts like *backwards rationalizability* (Penta (2015), Perea (2014)), *backward dominance* (Perea (2014)), *subgame perfect equilibrium* (Selten (1965)), *extensive-form perfect equilibrium* (Selten (1975)), *sequential equilibrium* (Kreps and Wilson (1982)) and *quasi-perfect equilibrium* (van Damme (1984)). According to these concepts, upon reaching an unexpected information set a player is free to believe that this is due to mistakes by his opponents in executing their planned strategies. Hence, a player is not required to try to learn from the past, even if doing so could refine his beliefs about the opponents' unobserved past, present and future moves.

In a sense, the forward induction concept of *strong rationalizability* (Pearce (1984), Battigalli (1997)), also known as *extensive-form rationalizability,* takes the other extreme by excluding, whenever possible, mistakes in the execution of planned strategies. However, to make this possible, a player may need to ascribe beliefs to the opponents that assume only a limited degree of rationality of their respective opponents. This, in turn, may lead to counterintuitive predicted behavior after surprises. The example in Figure 1, which is a variant of Figure 3 in Reny (1992), will illustrate this.

Upon reaching h_2, player 2 is forced to believe that player 1 chooses strategy (b, f), as this is the only way for player 1 to get at least 5 – a payoff he could guarantee by choosing a. At the same time, player 2 must believe that player 1 ascribes a high probability to player 2 behaving irrationally at h_4. The unique best reply for player 2 is to choose strategy (d, g). However, if player 1 in fact believes that player 2 will choose rationally at h_4, and chooses rationally himself in the remainder of the game, then player 1 would choose e at h_3, yielding an extremely low payoff for player 2.

We find the prediction (d, g) for player 2 unsatisfactory, because signs of irrationality by an opponent in the past – even if it is only by him ascribing low levels of rationality to the other players, and choosing optimally upon it – does not necessarily mean that this opponent will continue to do so in the future. Indeed, in the example above, how can player 2 be sure that player 1 will not "wake up" at h_3, but rather continue to hold the unreasonable belief that player 2 will choose irrationally

[1]It is often claimed that backward induction first appeared in Zermelo (1913) in the proof of his famous theorem on chess. However, Zermelo did not assume a stopping rule for chess, and hence the game he considered did not have a finite horizon. Therefore, he could not use backward induction.

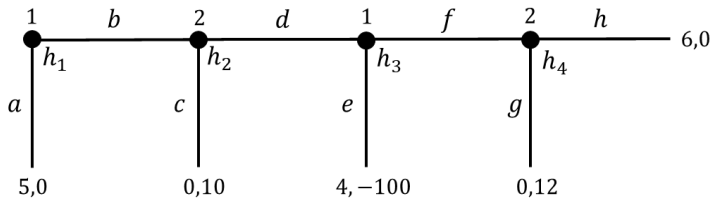

Figure 1: Strong rationalizability may lead to counterintuitive behavior

at h_4? It could as well be that player 1 held a reasonable belief about player 2 from the beginning, but chose b by mistake. As any good chess player knows, past mistakes by an opponent do not guarantee that he will make any mistakes in the future. In this sense, it is very dangerous to draw too strong conclusions from the opponents' past irrational behavior. But this is precisely what strong rationalizability does in many instances. As such, this concept is not robust with respect to players increasing their confidence in the opponents' future rationality after observing surprising moves.

Overall, we thus see some shortcomings with both backward and forward induction reasoning as discussed above. In response, we propose a rationalizability concept for dynamic games – different from both strong rationalizability and pure backward induction reasoning – which we call *forward and backward rationalizability.* In this new concept, we require players to learn from the past, but only as much as is consistent with "fully rational behavior" in the future.

More formally, the concept proceeds by applying the forward induction concept of strong rationalizability in a backward inductive fashion: We start by applying strong rationalizability to the last period in the game, which results in restrictions on strategies and beliefs there. Taking these restrictions as given, we then apply the strong rationalizability procedure to the parts of the game that start at the penultimate period, and so on, until we reach the beginning of the game.

We then ask: What does the new concept of forward and backward rationalizability entail in terms of reasoning? The epistemic characterization of strong rationalizability as given by Battigalli and Siniscalchi (2002) relies on *strong belief in rationality,* which means that a player, whenever possible, should believe that his opponents are choosing rationally in the whole game. On top of this, Battigalli and Siniscalchi (2002) require that a player must also believe, whenever possible, that his opponents choose rationally in the whole game while strongly believing in the other players' rationality. Iterating this argument leads to *rationality and common strong belief in rationality* – a set of epistemic conditions for which Battigalli and Siniscalchi (2002) show that they characterize strong rationalizability.

Recall that our concept imposes strong rationalizability from any period onwards in a backward inductive fashion, starting at the last period and then working backwards towards the first period.

It is shown in Theorem 4.1 that our concept can be characterized epistemically by (a) first imposing common strong belief in rationality from the last period onwards, (b) then imposing common strong belief in rationality from the penultimate period onwards, keeping the restrictions from (a), and so on, until we reach the beginning of the game.

As a consequence, when a player finds himself at an information set h he looks for the earliest period m before h such that some iterations of strong belief in rationality from period m onwards can explain the event of reaching h. In this case, the player will believe that *common* strong belief in rationality from period $m+1$ onwards will hold. This yields a best rationalization principle for the new concept. In particular, even at information sets a player did not expect to be reached, this player will always ascribe the highest possible degree of rationality, from the earliest possible information set onwards, to his opponents.

Intuitively, the reasoning process of a player choosing forward and backward rationalizable strategies can thus be described as follows: "Yesterday I believed that my opponents are rational from then (yesterday) on, believed that everybody else believed that everybody (else) is rational from then on, and so forth. If this is not falsified by today's observations, I should continue to believe not only that everybody else is rational from today on, and so forth, but also that everybody else *was rational from yesterday on*, and so forth. The same applies to the day before yesterday, the day before that, and so on."

We next investigate how the new concept relates to existing concepts, such as strong rationalizability and backwards rationalizability. In the example of Figure 1, for instance, the new concept coincides with backward induction. However, there are other games where the concept is different, in terms of strategies, from both pure backward induction reasoning and strong rationalizability. Consider, for instance, the game in Figure 2.

Strong rationalizability would reason as follows: At h_2, player 2 must believe that player 1 chooses strategy (a, e), as this is the only strategy reaching h_2 that would yield player 1 at least 2 – a payoff he could guarantee by choosing b at the beginning. But then, player 2 would choose d, and player 1, anticipating this, would choose b.

Our concept of forward and backward rationalizability proceeds differently: In the last subgame, at h_3, it imposes no restrictions. Now consider the subgame starting at h_2, which is the classical Battle of the Sexes game with an outside option for player 2. Our concept uniquely selects the forward induction strategies (c, h) and f in this subgame. Finally, we turn to the whole game. Given the earlier restrictions, player 1 must believe that player 2 will choose (c, h), and therefore will choose b himself. In particular, it predicts that player 2 will choose (c, h) and not d, as strong rationalizability predicts.

Our new concept thus yields a different strategy for player 2 than strong rationalizability, but it induces the same outcome – player 1 choosing b at the beginning. In Theorem 6.3 we show that this is no coincidence: The two concepts will always yield the same outcomes, but may differ in terms of strategies.

When compared to the pure backward induction concept of backwards rationalizability, our con-

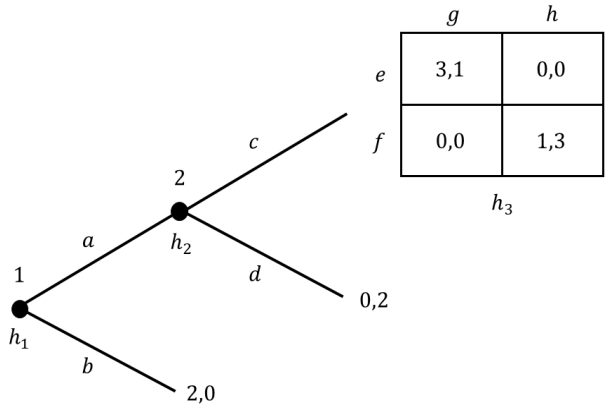

Figure 2: Battle of the sexes with double outside option

cept of forward and backward rationalizability always yields a refinement of the latter – not only in terms of outcomes but also in terms of strategies. See Theorem 5.1. In the game of Figure 2, for instance, backwards rationalizability yields the same strategies for player 1 as our concept, which is only strategy b, but allows for strategies (c, h) and d for player 2 whereas our concept only allows for (c, h). In fact, the new concept of forward and backward rationalizability can be viewed as a backward looking strengthening of the purely forward looking concept of backwards rationalizability: On top of always believing that the opponent will choose rationally in the future, which is what backwards rationalizability entails, we require a player to also explain (some of) the opponent's *past* choices whenever possible. In particular, after observing a surprising move by an opponent, the new concept requires the player to always reason in accordance with common belief in future rationality (Perea (2014)), and thus to assign the highest possible degree of *future* rationality to the opponent.

Although our concept is equivalent to strong rationalizability in terms of outcomes, we believe that strategies matter. Surprises and mistakes *do* happen regularly in reality, and therefore we find it important to critically analyze how players react to surprises. Indeed, a satisfactory theory of rational behavior should also describe how rational players act and reason upon observing surprising, or even irrational, behavior by their opponents. As we have argued in Figure 1, we think that strong rationalizability does not provide an appealing theory for how players react to such surprises. Different from strong rationalizability, under the concept of forward and backward rationalizability a player will never believe that an opponent will choose irrationally in the future, nor will a player attribute an unreasonable belief to an opponent. In that sense, we believe that the new concept provides a more satisfactory theory for how players react to surprises, and that it offers an alternative plausible rationale for strongly rationalizable outcomes.

An important immediate consequence of the two results mentioned above is that in every dynamic

game with imperfect information, every strongly rationalizable outcome is also induced by some profile of backwards rationalizable strategies. Catonini (2020) and Perea (2017) have already proven this result for dynamic games with observed past choices, but we show that this property even holds for games with imperfectly observed past choices. This may be viewed as a generalization of Battigalli's theorem, which states that in every dynamic game with perfect information and without relevant ties, the unique strongly rationalizable outcome is the backward induction outcome.

We finally investigate how players reason if they learn that the game was actually preceded by some earlier moves they were initially unaware of. Traditionally, we analyze a game by assuming that all players know that this is the game being played, and we may be interested in the players' behavior at "surprising" subgames, which were initially not expected to be reached. But instead of assuming that a player finds himself in a "surprising" subgame, it may also happen that a player initially views this subgame as the "whole game", and then learns that this game was actually preceded by some past moves.[2] If this happens, this could be a reason for the player to *refine* his reasoning, but, so we argue, it should never be a reason to completely *overthrow* the reasoning he did before. After all, the player reasoned himself towards a set of possible beliefs and strategies based on the accurate description of the game from now on, and the fact that the game actually started earlier does not change the accuracy of this description. If the game was in fact preceded by some past moves, this additional information should only lead to sharper predictions, but not to new beliefs and strategies which were considered irrational before receiving this new piece of information. Indeed, it is always possible that the moves preceding the game were chosen by mistake. This principle, that new information about past moves should only lead players to refine their reasoning, but not to overthrow their reasoning, is called *supergame monotonicity.*

We do not only see supergame monotonicity as a desideratum, but as a property that any solution concept for dynamic games based on rationality of the players should have. To keep players staying on a path that leads to an outcome consistent with the solution concept, players should also behave according to the solution concept once they are off the path. Otherwise, there is no guarantee that preventing the players from deviating is based on rational grounds.

This is exemplified by the distinction between Nash equilibrium and subgame perfect equilibrium in dynamic games. Nash equilibria in dynamic games that are not subgame perfect involve incredible threats that are inconsistent with Nash equilibria in unreached subgames. Therefore, it might actually be rational for some player to reach such a subgame. Hence, Nash equilibrium violates supergame monotonicity, while subgame perfect equilibrium satisfies it.

As is easily seen, the game of Figure 1 shows that strong rationalizability violates supergame monotonicity. If the game were to start at h_2, then strong rationalizability would uniquely select the strategy c for player 2. However, if player 2 learns that the game in fact started at h_1, then strong rationalizability would uniquely select strategy (d, g) for player 2, and therefore the previous reasoning of player 2 would be completely overthrown by this new information. In contrast, the pure

[2] This would be a special instance of an extensive-form game with unawareness. See, for example, Heifetz, Meier and Schipper (2013).

backward induction concept of backwards rationalizability *does* satisfy supergame monotonicity.

We show in Theorem 8.1 that also the new concept of forward and backward rationalizability always satisfies supergame monotonicity. Consider, for instance, the game from Figure 2. If the game were to start at h_3, our concept selects both e and f for player 1. However, if player 1 learns that the game started at h_2, this additional information will refine the set of possible choices for player 1 to only f.

The paper is organized as follows: In Section 2 we lay out the basic definitions and notation for dynamic games. In Section 3 we formally define the strong rationalizability procedure, the backwards rationalizability procedure and the new forward and backward rationalizability procedure, and prove the existence of the latter concept. In Section 4 we provide an epistemic characterization of our concept. In Section 5 we show that the new concept constitutes a refinement, in terms of strategies, of backwards rationalizability, whereas we prove in Section 6 that in terms of outcomes it is equivalent to strong rationalizabity. In Section 7 we show how these results imply a generalization of Battigalli's theorem, by proving that in every dynamic game with possibly imperfect information, all strongly rationalizable outcomes are also backwards rationalizable. In Section 8 we discuss the principle of supergame monotonicity. The appendix contains all the proofs.

2 Preliminaries

In this section we introduce our model of dynamic games and establish the notation we use. Subsequently, we define strategies, conditional belief vectors, strong belief, and optimal choice at information sets.

2.1 Dynamic Games

In this paper we consider finite dynamic games that allow for simultaneous moves and imperfect information, and where every action and history can be indexed by time. Formally, a *dynamic game* is a tuple $\Gamma = (I, T, (P^m)_{m \in T}, I^a, (A_i, H_i)_{i \in I}, Z, (u_i)_{i \in I})$, where

(a) I is the finite set of *players*;

(b) $T = \{L, L+1, \ldots, M+1\}$ is the collection of *time periods*, with $L, M \in \mathbf{Z}$ and $L \leq M$. The reason we allow the first time period to be different from 1 is that later we look at subgames and supergames of Γ, which are games by themselves and may start at a later or earlier time period;

(c) for every period $m \in T$ we denote by P^m the set of *past action profiles*, or *histories*, at period m, to be defined below. By $P := \cup_{m \in T} P^m$ we denote the set of all possible histories;

(d) the mapping I^a assigns to every history $p \in P$ the set of *active players* $I^a(p) \subseteq I$ who must choose after history p. If $I^a(p)$ contains more than one player, there are simultaneous moves after p. We also allow $I^a(p)$ to be empty, in which case no player chooses after history p. In this case, p will be a *terminal* history. By P_i we denote the set of histories $p \in P$ with $i \in I^a(p)$;

(e) for every player i, the mapping A_i assigns to every history $p \in P_i$ the finite set of *actions* $A_i(p)$ from which player i can choose after history p;

(f) for every period $m \in T$, the set P^m of histories at period m is inductively defined by $P^L := \{p^L\}$, and

$$P^m := \{(p^{m-1}, (a_i)_{i \in I^a(p^{m-1})}) \mid p^{m-1} \in P^{m-1} \text{ and } a_i \in A_i(p^{m-1}) \text{ for all } i \in I^a(p^{m-1})\}$$

for every $m \in \{L+1, \ldots, M+1\}$. That is, a history in P^m describes which actions have been chosen before period m. We assume that $I^a(p) = \emptyset$ for every history $p \in P^{M+1}$. That is, every history in the last period $M+1$ will be a terminal history. We say that a history p *precedes* a history p' (or, p' *follows* p) if p' results by adding some actions after p;

(g) for every player i, the mapping H_i assigns to every period $m \in \{L, ..., M\}$ a partition $H_i^m = \{h_i^{m.1}, ..., h_i^{m.K_{im}}\}$ of the set of histories $P_i \cap P^m$ into *information sets.* The interpretation is that for every information set $h \in H_i^m$, player i cannot distinguish between any two histories $p, p' \in h$. However, the player always knows in which time period he is. By $H_i := \cup_{m \in T} H_i^m$ we denote the collection of information sets for player i, whereas $H := \cup_{i \in I} H_i$ denotes the collection of all information sets of all players. For a given period m, we denote by $H^m := \cup_{i \in I} H_i^m$ the collection of all information sets at period m. As player i's set of actions must be the same for every two histories between which he cannot distinguish, we require, for every player i, and every information set $h \in H_i$, that $A_i(p) = A_i(p')$ for every $p, p' \in h$. By $A_i(h)$ we denote the unique set of actions between which player i can choose at information set $h \in H_i$. We require that $A_i(h) \cap A_i(h') = \emptyset$ for every two distinct information sets $h, h' \in H_i$. We also assume *perfect recall,* which means that player i always remembers which actions he chose in the past, and which information he had about the opponents' past actions at previous periods. Formally, for every player i, every information set $h \in H_i$, and every two histories $p, p' \in H_i$, the set of player i actions in p and p' must be the same (and consequently, the collection of player i information sets that p and p' cross must be the same). For every two information sets $h, h' \in H$, we say that h precedes h' (or h' follows h) if there is a history $p \in h$ and a history $p' \in h'$ such that p precedes p';

(h) Z is the set of *terminal histories* given by

$$Z := \{p \in P \mid I^a(p) = \emptyset\}.$$

These are thus the histories after which no player makes a move. As such, a terminal history $z \in Z$ describes which actions have been chosen from the beginning to the end;

(i) for every player i, the *utility function* $u_i : Z \to \mathbf{R}$ assigns to every terminal history $z \in Z$ some utility $u_i(z)$.

2.2 Strategies

A strategy for player i assigns an available action to every information set at which player i is active, and that is not excluded by earlier actions in the strategy. Formally, let $\tilde{s}_i$ be a mapping that

assigns to *every* information set $h \in H_i$ some action $\tilde{s}_i(h) \in A_i(h)$. We call $\tilde{s}_i$ a *complete strategy.* Then, a history $p \in P$ is *excluded* by $\tilde{s}_i$ if there is some information set $h \in H_i$, with some history $p' \in h$ preceding p, such that $\tilde{s}_i(h)$ is different from the unique player i action at p' leading to p. An information set $h \in H$ is excluded by $\tilde{s}_i$ if all histories in h are excluded by $\tilde{s}_i$. The *strategy* induced by $\tilde{s}_i$ is the restriction of $\tilde{s}_i$ to those information sets in H_i that are not excluded by $\tilde{s}_i$. A mapping $s_i : \tilde{H}_i \to \cup_{h \in \tilde{H}_i} A_i(h)$, where $\tilde{H}_i \subseteq H_i$, is a *strategy* for player i if it is the strategy induced by a complete strategy.[3] By S_i we denote the set of strategies for player i, and by $S_{-i} := \times_{j \neq i} S_j$ the set of strategy combinations for i's opponents.

Consider a strategy profile $s = (s_i)_{i \in I}$ in $\times_{i \in I} S_i$. Then, s induces a unique terminal history $z(s)$. We say that the strategy profile s *reaches* a history p if p precedes $z(s)$. Similarly, the strategy profile s is said to reach an information set h if s reaches a history in h.

For a given player i and information set $h \in H_i$, we define the sets

$$\begin{aligned} S(h) &: = \{s \in \times_{i \in I} S_i \mid s \text{ reaches } h\}, \\ S_i(h) &: = \{s_i \in S_i \mid \text{there is some } s_{-i} \in S_{-i} \text{ such that } (s_i, s_{-i}) \in S(h)\}, \text{ and} \\ S_{-i}(h) &: = \{s_{-i} \in S_{-i} \mid \text{there is some } s_i \in S_i \text{ such that } (s_i, s_{-i}) \in S(h)\}. \end{aligned}$$

Intuively, $S_i(h)$ is the set of strategies for player i that allow for information set h to be reached, whereas $S_{-i}(h)$ is the set of opponents' strategy combinations that allow for h to be reached. By perfect recall it holds, for every player i and every information set $h \in H_i$, that $S(h) = S_i(h) \times S_{-i}(h)$. For a given strategy $s_i \in S_i$, we denote by $H_i(s_i) := \{h \in H_i \mid s_i \in S_i(h)\}$ the collection of information sets for player i that the strategy s_i allows to be reached.

2.3 Conditional Belief Vectors and Strong Belief

For a finite set X, let $\Delta(X)$ be the set of probability distributions on X. A *conditional belief vector* for player i is a mapping b_i that assigns to every information set $h \in H_i$ a probabilistic belief $b_i(h) \in \Delta(S_{-i}(h))$ about the opponents' strategy combinations that allow for h to be reached. Moreover, a conditional belief vector b_i is required to satisfy *Bayesian updating* whenever possible. That is, for every $h, h' \in H_i$ where h precedes h' and $b_i(h)(S_{-i}(h')) > 0$ we have that

$$b_i(h')(s_{-i}) = \frac{b_i(h)(s_{-i})}{b_i(h)(S_{-i}(h'))}$$

for every $s_{-i} \in S_{-i}(h')$.[4] Let B_i be the set of all conditional belief vectors for player i that satisfy Bayesian updating whenever possible.

[3] What we call a "strategy" is sometimes called a "plan of action" in the literature (Rubinstein (1991)), and what we call a "complete strategy" is often called a "strategy". The advantage of the notion of "strategy" we use here, is that it reduces the description of strategies and strategy sets when carrying out any of the rationalizability procedures.

[4] By abuse of notation, we write $b_i(h)(s_{-i})$ instead of $b_i(h)(\{s_{-i}\})$.

For a given player i, consider a set of opponents' strategy combinations $D_{-i} \subseteq S_{-i}$. Say that a conditional belief vector $b_i \in B_i$ *strongly believes* D_{-i} if for every information set $h \in H_i$ with $S_{-i}(h) \cap D_{-i} \neq \emptyset$ we have that $b_i(h)(D_{-i}) = 1$.[5] That is, b_i assigns full probability to strategy combinations in D_{-i} whenever this is possible.

2.4 Optimal Choice at Information Sets

Recall that every strategy combination $s = (s_i)_{i \in I}$ induces a terminal history $z(s)$. For a strategy s_i, an information set $h \in H_i(s_i)$ and a conditional belief vector b_i, let

$$u_i(s_i, b_i(h)) := \sum_{s_{-i} \in S_{-i}(h)} b_i(h)(s_{-i}) \cdot u_i(z(s_i, s_{-i}))$$

be the *expected utility* induced by strategy s_i at information h under the conditional belief vector b_i. A strategy s_i is *optimal* for b_i at $h \in H_i(s_i)$ if

$$u_i(s_i, b_i(h)) \geq u_i(s_i', b_i(h)) \text{ for all } s_i' \in S_i(h).$$

For a given time period m, strategy s_i is said to be *optimal* for b_i *from period m onwards* if for every period $\tau \geq m$ and every information set $h \in H_i(s_i) \cap H_i^\tau$, strategy s_i is optimal for b_i at h.

Note that if a strategy s_i does not allow any information set in H^τ with $\tau \geq m$ to be reached then, by definition, s_i is (vacuously) optimal from period m onwards for every $b_i \in B_i$. It can be shown that the following is true:

Remark 2.1 *For every conditional belief vector $b_i \in B_i$, every period $m \in \{L, ..., M\}$ and every information set $h \in H^m$, there is a strategy s_i that allows h to be reached and that is optimal for b_i from period m onwards.*

3 Definition of the Concepts

In this section we recall the concepts of *strong rationalizability* and *backwards rationalizability,* and introduce the new concept of *forward and backward rationalizability.*

3.1 Strong Rationalizability

The *strong rationalizability* procedure (Pearce (1984), Battigalli (1997)) is a forward induction procedure that recursively eliminates strategies and conditional belief vectors for every player. The main idea is that a player, whenever possible, must believe that his opponents are implementing strategies

[5] Here, we use the convention that $b_i(h)(D_{-i}) := b_i(h)(D_{-i} \cap S_{-i}(h))$.

that are optimal from the beginning (that is, from period L onwards). On top of this, a player must also believe, whenever possible, that his opponents are not only implementing strategies that are optimal from the beginning, but also that the opponents engage in the reasoning described above. Iterating these reasoning steps then leads to strong rationalizability.

Definition 3.1 (Strong rationalizability) ***Round*** 0***:*** *For every player i, set $B_i^0 := B_i$ and $S_i^0 := S_i$.*

Round $k \geq 1$***:*** *For every player i, set*

$$B_i^k := \{b_i \in B_i^{k-1} \mid b_i \text{ strongly believes } S_{-i}^{k-1}\}$$

and

$$S_i^k := \{s_i \in S_i^{k-1} \mid s_i \text{ is optimal for some } b_i \in B_i^k \text{ from period } L \text{ onwards}\}.$$

Strategy $s_i \in S_i$ is ***strongly rationalizable*** *if $s_i \in S_i^k$ for all $k \geq 0$. Conditional belief vector b_i is* ***strongly rationalizable*** *if $b_i \in B_i^k$ for all $k \geq 0$.*

We will see that the new concept of forward and backward rationalizability applies the strong rationalizability procedure in a backward inductive fashion.

3.2 Backwards Rationalizability

The concept of *backwards rationalizability* (Perea (2014), Penta (2015)) is purely forward looking, and can thus be viewed as a genuine backward induction concept. This can also be seen from the fact that the resulting strategies can be characterized by *common belief in future rationality* (Perea (2014)), stating that a player always believes that his opponents will choose rationally now and in the future, always believes that his opponents always believe that the other players will choose rationally now and in the future, and so on.

Definition 3.2 (Backwards rationalizability) *Consider a dynamic game with time periods $L, ..., M+1$.*

Period M, ***round*** 0. *Set $S_i^{M.0} := S_i$ and $B_i^{M.0} := B_i$ for all players i.*

Period M, ***round*** $k \geq 1$. *For every player i, set*

$$B_i^{M.k} := \{b_i \in B_i^{M.k-1} \mid b_i(h)(S_{-i}^{M.k-1}) = 1 \text{ for all } h \in H_i^M\},$$

and

$$S_i^{M.k} := \{s_i \in S_i^{M.k-1} \mid s_i \text{ is optimal for some } b_i \in B_i^{M.k} \text{ from period } M \text{ onwards}\}.$$

Continue until $B_i^{M.K_M} = B_i^{M.K_M+1}$ and $S_i^{M.K_M} = S_i^{M.K_M+1}$ for some round K_M, and for all players i.

Period $m \leq M-1$, ***round*** 0. *Set* $S_i^{m.0} := S_i^{m+1.K_{m+1}}$ *and* $B_i^{m.0} := B_i^{m+1.K_{m+1}}$ *for all players* i.

Period $m \leq M-1$, ***round*** $k \geq 1$. *For every player* i, *let*

$$B_i^{m.k} := \{b_i \in B_i^{m.k-1} \mid b_i(h)(S_{-i}^{m.k-1}) = 1 \text{ for all } h \in H_i^m\},$$

and

$$S_i^{m.k} := \{s_i \in S_i^{m.k-1} \mid s_i \text{ is optimal for some } b_i \in B_i^{m.k} \text{ from period } m \text{ onwards}\}.$$

Continue until $B_i^{m.K_m} = B_i^{m.K_m+1}$ *and* $S_i^{m.K_m} = S_i^{m.K_m+1}$ *for some round* K_m, *and for all players* i.

A strategy s_i *is* ***backwards rationalizable*** *if* $s_i \in S_i^{L.K_L}$, *and a conditional belief vector* b_i *is* ***backwards rationalizable*** *if* $b_i \in B_i^{L.K_L}$.

The definition we have provided here uses the backwards order of elimination. That is, we start by eliminating at the ultimate period, subsequently we do the eliminations at the penultimate period, and so on, until we reach the beginning of the game. Perea (2014)'s definition is different, as in every round it (potentially) performs eliminations at each of the information sets in each of the periods. However, it is argued in Section 6.3 of Perea (2014) that the definition we provide here is equivalent, in terms of strategies and conditional belief vectors selected, to the definition in Perea (2014).

3.3 Forward and Backward Rationalizability

The new concept of *forward and backward rationalizability* combines elements from the strong rationalizability procedure and the backwards rationalizability procedure. Like the backwards rationalizability procedure, it proceeds in a backward inductive fashion by first performing eliminations at the ultimate period, and then proceeding backwards until we reach the beginning of the game. However, when we reach a certain period m in this way then, in line with strong rationalizability, we also require players at later periods to reason about the opponents' past moves at period m. This is fundamentally different from backwards rationalizability, where players at a given period are only required to reason about the opponents' moves at this period and *future* periods.

Definition 3.3 (Forward and backward rationalizability) *Consider a dynamic game with time periods* $L, ..., M+1$.

Period M, ***round*** 0. *Set* $S_i^{M.0} := S_i$ *and* $B_i^{M.0} := B_i$ *for all players* i.

Period M, ***round*** $k \geq 1$. *For every player* i, *set*

$$B_i^{M.k} := \{b_i \in B_i^{M.k-1} \mid b_i \text{ strongly believes } S_{-i}^{M.k-1}\},$$

and

$$S_i^{M.k} := \{s_i \in S_i^{M.k-1} \mid s_i \text{ is optimal for some } b_i \in B_i^{M.k} \text{ from period } M \text{ onwards}\}.$$

Continue until $B_i^{M.K_M} = B_i^{M.K_M+1}$ *and* $S_i^{M.K_M} = S_i^{M.K_M+1}$ *for some round* K_M*, and for all players* i*.*

Period $m \leq M-1$, ***round*** 0*. Set* $S_i^{m.0} := S_i^{m+1.K_{m+1}}$ *and* $B_i^{m.0} := B_i^{m+1.K_{m+1}}$ *for all players* i*.*

Period $m \leq M-1$, ***round*** $k \geq 1$*. For every player* i*, let*

$$B_i^{m.k} := \{b_i \in B_i^{m.k-1} \mid b_i \textit{ strongly believes } S_{-i}^{m.k-1}\},$$

and

$$S_i^{m.k} := \{s_i \in S_i^{m.k-1} \mid s_i \textit{ is optimal for some } b_i \in B_i^{m.k} \textit{ from period } m \textit{ onwards}\}.$$

Continue until $B_i^{m.K_m} = B_i^{m.K_m+1}$ *and* $S_i^{m.K_m} = S_i^{m.K_m+1}$ *for some round* K_m*, and for all players* i*.*

A strategy s_i *is* ***forward and backward rationalizable*** *if* $s_i \in S_i^{L.K_L}$*, and a conditional belief vector* b_i *is* ***forward and backward rationalizable*** *if* $b_i \in B_i^{L.K_L}$*.*

In the initial step we thus apply the strong rationalizability procedure to those parts of the game that start at the final decision period M, where every player makes at most one choice. In the next step, we turn to the parts of the game that start at period $M-1$, and apply the strong rationalizability procedure there, taking as an input the restrictions from the initial step. The transition from period m to $m-1$ works in an analogous way, by taking as input the restrictions from period m. We continue in this fashion until we reach the beginning of the game. Hence, we apply the strong rationalizability procedure in a backward inductive fashion.

For a given player i and information set h in period m, let $\tau \leq m$ be the earliest period such that $S_i^{\tau.k} \cap S_i(h) \neq \emptyset$ for some round $k \geq 0$. For this period τ, let l be the highest such round with $S_i^{\tau.l} \cap S_i(h) \neq \emptyset$.[6] Then, we denote by $S_i^{fbr}(h) := S_i^{\tau.l} \cap S_i(h)$ the set of strategies that is predicted conditional on reaching information set h.

Note that according to our concept, a player $j \neq i$ with $h \in H_j$ may believe at information set h that player i chooses a strategy which is not in $S_i^{fbr}(h)$ in case there are more than two players. Indeed, it may be that $S_i^{fbr}(h) = S_i^{\tau.l} \cap S_i(h)$, whereas for a third player $n \neq i, j$ we have that $S_n^{fbr}(h) = S_n^{\tau'.l'} \cap S_n(h)$, where either $\tau' > \tau$ or ($\tau' = \tau$ and $l' < l$). In that case, a conditional belief vector b_j that is forward and backward rationalizable may at h assign positive probability to opponents' strategy combinations containing some $s_i \in S_i^{\tau'.l'} \setminus S_i^{\tau.l}$.

By construction, our procedure will refine the conditional beliefs of player i at some information set $h \in H_i^m$ until we reach a period $\tau < m$ where $S_{-i}^{\tau.k} \cap S_{-i}(h)$ becomes empty for some $k \geq 1$. In that case, player i's conditional beliefs at h will thus be anchored at period τ and round $k-1$. This implies that at h, player i believes that his opponents were "level $(k-1)$-rational" from period τ onwards, and "fully rational" from period $\tau+1$ onwards.

It is not difficult to prove that the concept always yields at least one strategy and conditional belief vector for every player.

[6] If $m > L$ then $\tau < m$, since h is always reachable by a strategy in $S_i^{m-1.0} = S_i^{m.K_m}$.

Theorem 3.1 (Existence) *For every player there are always at least one strategy and one conditional belief vector that are forward and backward rationalizable.*

Thus, it can never happen that all remaining strategies or conditional belief vectors for a given player are eliminated at a particular round.

3.4 Examples

We will now illustrate the forward and backward rationalizability procedure by means of two examples. The first example is rather small, and is simply there to illustrate how the procedure works. The second example is large, and shows how the procedure can still be applied efficiently if there are many information sets and periods in the game.

Example 1. Consider first the example from Figure 2 in the introduction. There are four time periods in this game: Period 1 with information set h_1, period 2 with information set h_2, period 3 with information set h_3, and period 4 where the game ends. That is, $L = 1$ and $M = 3$. We will now run the forward and backward rationalizability procedure, starting at period 3.

Period 3. We have that $B_1^{3.1} = B_1$ and $B_2^{3.1} = B_2$. For player 1, both strategies (a, e) and (a, f) are optimal from period 3 onwards for some conditional belief vector in $B_1^{3.1}$, and similarly for player 2's strategies (c, g) and (c, h). Note that the strategies b and d are vacuously optimal from period 3 onwards for some conditional belief vector in $B_1^{3.1}$ and $B_2^{3.1}$, respectively. Thus,

$$S_1^{3.1} = S_1 = \{b, (a, e), (a, f)\} \text{ and } S_2^{3.1} = S_2 = \{d, (c, g), (c, h)\},$$

and this is where the procedure at Period 3 terminates.

Period 2. *Round 1.* We have that $B_1^{2.1} = B_1$ and $B_2^{2.1} = B_2$. For player 2, strategy (c, g) is not optimal from period 2 onwards for any conditional belief vector in $B_2^{2.1}$. In turn, strategies (c, h) and d *are* optimal from period 2 onwards for some conditional belief vector in $B_2^{2.1}$. Thus,

$$S_2^{2.1} = \{d, (c, h)\}.$$

Round 2. We then have

$$B_1^{2.2} = \{b_1 \in B_1 \mid b_1(h_1)(\{d, (c, h)\}) = 1 \text{ and } b_1(h_3)((c, h)) = 1\}.$$

Since for player 1 only strategies b and (a, f) are optimal from period 2 onwards for some conditional belief vector in $B_1^{2.2}$, it follows that

$$S_1^{2.2} = \{b, (a, f)\}.$$

Round 3. This implies that

$$B_2^{2.3} = \{b_2 \in B_2 \mid b_2(h_2)((a, f)) = 1 \text{ and } b_2(h_3)((a, f)) = 1\}.$$

For player 2, only strategy (c, h) is optimal from period 2 onwards for some conditional belief vector in $B_2^{2.3}$, and we thus conclude that

$$S_2^{2.3} = \{(c, h)\}.$$

Round 4. We then have that

$$B_1^{2.4} = \{b_1 \in B_1 \mid b_1(h_1)((c, h)) = 1 \text{ and } b_1(h_3)((c, h)) = 1\},$$

after which no further eliminations are possible in period 2.

Period 1. We start with the restrictions on the strategies and conditional belief vectors inherited from period 2. That is,

$$S_1^{1.0} = \{b, (a, f)\},\ B_1^{1.0} = \{b_1 \in B_1 \mid b_1(h_1)((c, h)) = 1 \text{ and } b_1(h_3)((c, h)) = 1\},$$
$$S_2^{1.0} = \{(c, h)\} \text{ and } B_2^{1.0} = \{b_2 \in B_2 \mid b_2(h_2)((a, f)) = 1 \text{ and } b_2(h_3)((a, f)) = 1\}.$$

For player 1, the only strategy that is optimal from period 1 onwards for some conditional belief vector in $B_1^{1.0}$ is b. We therefore have

$$S_1^{1.1} = \{b\}.$$

Afterwards, no further eliminations are possible. We thus conclude that the strategies selected by the forward and backward rationalizability procedure are b for player 1 and (c, h) for player 2.

On the other hand, as we have seen in the introduction, strong rationalizability selects the strategies b for player 1 and d for player 2. The intuition for this difference is the following: According to forward and backward rationalizability, player 2 asks at h_2: What is the earliest period m such that player 1's past behavior – that is, player 1 choosing a – can be explained by "full rationality" from period m onwards? This must be period 2. Indeed, from period 2 onwards, player 1 expects player 2 to choose (c, h), which makes it optimal for player 1, from period 2 onwards, to choose (a, f). In turn, if player 2 expects player 1 to choose (a, f), then it is optimal for player 2 to choose (c, h). This is a plausible theory for the reasoning and play from period 2 onwards.

However, if player 1 anticipates player 2 choosing (c, h), then it can never be optimal for player 1 to choose a at h_1. As such, player 1 choosing a cannot be explained by "full rationality" from period 1 onwards.

According to strong rationalizability, player 2 asks at h_2: Is there a strategy for player 1 involving his observed past move a that is optimal for *some* belief, even if this belief attributes irrational future strategies to player 2? This reasoning leads player 2 to believe that player 1 chooses (a, e), as this is the only strategy involving a that can possibly yield him at least 2. As a consequence, player 2 will choose d. Note, however, that stratey (a, e) can only yield player 1 at least 2 if he believes that player 2 irrationally chooses the strategy (c, g) in the future. As such, believing that player 1 chooses (a, e) cannot be part of a "fully rational" theory from period 1 onwards. Therefore, our concept discards this type of reasoning by player 2.

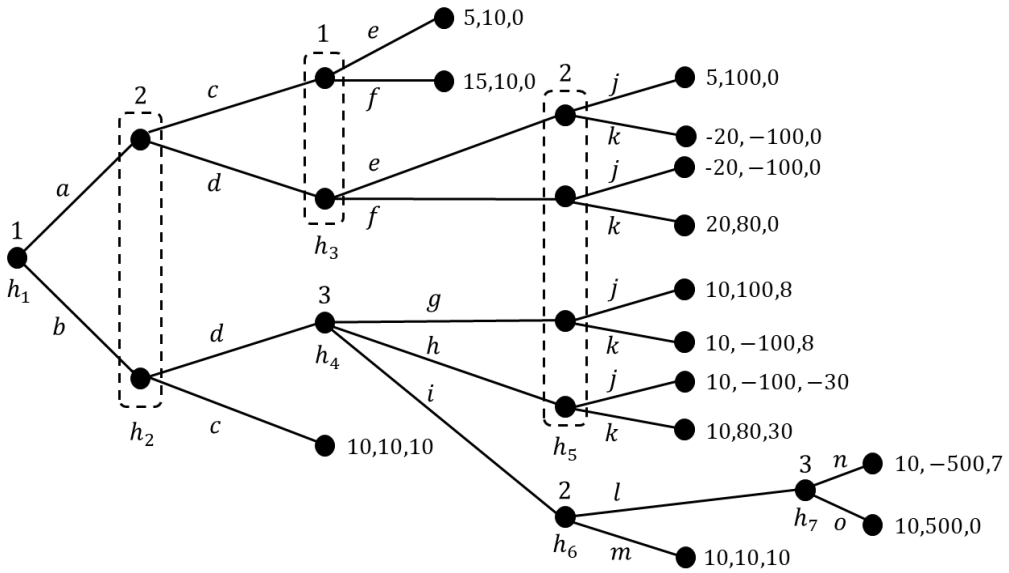

Figure 3: A dynamic game with non-trivial information sets

Example 2. We next move to a more complex dynamic game with "non-trivial" information sets, containing more than one history. Consider the dynamic game from Figure 3. Note that the information sets h_2, h_3 and h_5 are non-trivial.

In the dynamic game there are six time periods: Period 1 with information set h_1, period 2 with information set h_2, period 3 with information sets h_3 and h_4, period 4 with information sets h_5 and h_6, period 5 with information set h_7, and period 6 with no information sets but only terminal histories. To run the forward and backward rationalizability procedure, we thus start at period 5.

Period 5. At information set h_7, player 3's strategy (i, o) is never optimal for any conditional belief, whereas the strategy (i, n) is. Thus, only player 3's strategies g, h and (i, n) are optimal for some conditional belief vector from period 5 onwards. As such,

$$S_3^{5.1} = \{g, h, (i, n)\}.$$

Moreover, we have that $S_1^{5.1} = S_1$ and $S_2^{5.1} = S_2$. Hence,

$$B_1^{5.2} = \{b_1 \in B_1 \mid b_1 \text{ strongly believes } S_2 \times \{g, h, (i, n)\}\}$$
$$= \{b_1 \in B_1 \mid b_1(h_1)(S_2 \times \{g, h, (i, n)\}) = b_1(h_3)(S_2 \times \{g, h, (i, n)\}) = 1\},$$

and

$$B_2^{5.2} = \{b_2 \in B_2 \mid b_2 \text{ strongly believes } S_1 \times \{g, h, (i, n)\}\}$$
$$= \{b_2 \in B_2 \mid b_2(h_2)(S_1 \times \{g, h, (i, n)\}) =$$
$$b_2(h_5)(S_1 \times \{g, h, (i, n)\}) = b_2(h_6)(\{b\} \times \{(i, n)\}) = 1\}.$$

Finally, $B_3^{5.2} = B_3$. As no further restrictions can be derived after this step, the procedure from Period 5 onwards is hereby complete.

Period 4. At h_5, both choices j and k of player 2 are optimal for some conditional belief vector in $B_2^{5.2}$.

At h_6, player 2's conditional belief must be part of $B_2^{5.2}$ above, and hence player 2 must at h_6 assign probability 1 to player 3 choosing n at h_7. But then, player 2's choice l cannot be optimal at h_6 for any conditional belief vector in $B_2^{5.2}$, whereas the choice m can. Hence,

$$S_2^{4.1} = \{(c, (d, j, m), (d, k, m)\}.$$

Moreover, $S_1^{4.1} = S_1$ and $S_3^{4.1} = S_3^{5.1} = \{g, h, (i, n)\}$. We then have that

$$\begin{aligned} B_1^{4.2} &= \{b_1 \in B_1^{5.2} \mid b_1 \text{ strongly believes } \{(c, (d, j, m), (d, k, m)\} \times \{g, h, (i, n)\}\} \\ &= \{b_1 \in B_1 \mid b_1(h_1)(\{(c, (d, j, m), (d, k, m)\} \times \{g, h, (i, n)\}) \\ &= b_1(h_3)(\{(c, (d, j, m), (d, k, m)\} \times \{g, h, (i, n)\}) = 1\} \end{aligned}$$

and

$$\begin{aligned} B_3^{4.2} &= \{b_3 \in B_3^{5.2} \mid b_3 \text{ strongly believes } S_1 \times \{(c, (d, j, m), (d, k, m)\}\} \\ &= \{b_3 \in B_3 \mid b_3(h_4)(\{b\} \times \{(d, j, m), (d, k, m)\} = 1\}, \end{aligned}$$

whereas $B_2^{4.2} = B_2^{5.2}$.

As no further restrictions can be derived after this step, the procedure from Period 4 onwards is hereby complete.

Period 3. At h_3, both of player 1's strategies (a, e) and (a, f) can be optimal for some conditional belief vector in $B_1^{4.2}$. Thus,

$$S_1^{3.1} = S_1.$$

At h_4, player 3's conditional belief must be part of $B_3^{4.2}$, and hence player 3 must believe that player 2 will choose m at h_6. As such, player 3 expects the utility 10 by choosing i at h_4, whereas g gives him only 8. This renders the strategy g suboptimal for player 3 at h_4. Hence,

$$S_3^{3.1} = \{h, (i, n)\}.$$

As player 2 is not active at any information in Period 3, we have that

$$S_2^{3.1} = S_2^{4.1} = \{(c, (d, j, m), (d, k, m)\}.$$

Thus,

$$\begin{aligned} B_1^{3.2} &= \{b_1 \in B_1^{4.2} \mid b_1 \text{ strongly believes } \{(c, (d, j, m), (d, k, m)\} \times \{h, (i, n)\}\} \\ &= \{b_1 \in B_1 \mid b_1(h_1)(\{(c, (d, j, m), (d, k, m)\} \times \{h, (i, n)\}) \\ &= b_1(h_3)(\{(c, (d, j, m), (d, k, m)\} \times \{h, (i, n)\}) = 1\}, \end{aligned}$$

and

$$B_2^{3.2} = \{b_2 \in B_2^{4.2} \mid b_2 \text{ strongly believes } S_1 \times \{h, (i,n)\}$$
$$= \{b_2 \in B_2 \mid b_2(h_2)(S_1 \times \{h, (i,n)\}) =$$
$$b_2(h_5)((\{(a,e),(a,f)\} \times \{h, (i,n)\}) \cup (\{b\} \times \{h\}))$$
$$= b_2(h_6)(\{b\} \times \{(i,n)\}) = 1\},$$

whereas

$$B_3^{3.2} = \{b_3 \in B_3^{4.2} \mid b_3 \text{ strongly believes } S_1 \times \{(c,(d,j,m),(d,k,m)\}\}$$
$$= \{b_3 \in B_3 \mid b_3(h_4)(\{b\} \times \{(d,j,m),(d,k,m)\} = 1\}.$$

Since we can derive no further restrictions after this step, this concludes the procedure from Period 3 onwards.

Period 2. At the information set h_2, each of player 2's strategies in $S_2^{3.1}$ can be optimal for some conditional belief vector in $B_2^{3.2}$. Hence, $S_2^{2.1} = S_2^{3.1}$. As a consequence, the strategy sets and the sets of conditional belief vectors for each of the players remain the same as in Period 3.

Period 1. *Round 1.* At h_1, player 1 can guarantee utility 10 by choosing b. Since the strategy (a,e) yields him at most 5, we conclude that the strategy (a,e) is suboptimal for player 1 at h_1. Therefore,

$$S_1^{1.1} = \{(a,f), b\}.$$

Moreover,

$$S_2^{1.1} = S_2^{3.1} = \{(c,(d,j,m),(d,k,m)\}$$

and

$$S_3^{1.1} = S_3^{3.1} = \{h, (i,n)\}.$$

Round 2. Thus, $B_1^{1.2} = B_1^{3.2}$, and

$$B_2^{1.2} = \{b_2 \in B_2^{3.2} \mid b_2 \text{ strongly believes } \{(a,f), b\} \times \{h, (i,n)\}$$
$$= \{b_2 \in B_2 \mid b_2(h_2)(\{(a,f), b\} \times \{h, (i,n)\}) =$$
$$b_2(h_5)((\{(a,f)\} \times \{h, (i,n)\}) \cup (\{b\} \times \{h\}))$$
$$= b_2(h_6)(\{b\} \times \{(i,n)\}) = 1\},$$

whereas

$$B_3^{1.2} = \{b_3 \in B_3^{3.2} \mid b_3 \text{ strongly believes } \{(a,f), b\} \times \{(c,(d,j,m),(d,k,m)\}\}$$
$$= \{b_3 \in B_3 \mid b_3(h_4)(\{b\} \times \{(d,j,m),(d,k,m)\} = 1\}.$$

Note that at h_5, player 2 can only assign positive probability to the opponents' strategy combinations in $(\{(a,f)\} \times \{h,(i,n)\}) \cup (\{b\} \times \{h\})$. Thus, at h_5 player 2 can only assign positive probability to the second and fourth history. But then, player 2 should definitely choose k at h_5, and we thus have that

$$S_2^{1.2} = \{(c,(d,k,m)\}.$$

Moreover, $S_1^{1.2} = S_1^{1.1}$ and $S_3^{1.2} = S_3^{1.1}$.

Round 3. As such,

$$\begin{gathered} B_1^{1.3} = \{b_1 \in B_1^{1.2} \mid b_1 \text{ strongly believes } \{(c,(d,k,m)\} \times \{h,(i,n)\}\} \\ = \{b_1 \in B_1 \mid b_1(h_1)(\{(c,(d,k,m)\} \times \{h,(i,n)\}) = b_1(h_3)(\{(c,(d,k,m)\} \times \{h,(i,n)\}) = 1\} \end{gathered}$$

and

$$\begin{gathered} B_3^{1.3} = \{b_3 \in B_3^{1.2} \mid b_3 \text{ strongly believes } \{(a,f),b\} \times \{(c,(d,k,m)\}\} \\ = \{b_3 \in B_3 \mid b_3(h_4)(\{b\} \times \{(d,k,m)\} = 1\}, \end{gathered}$$

whereas $B_2^{1.3} = B_2^{1.2}$.

Hence, at h_1 player 1 believes that player 2 chooses either c or (d,k,m). But then, by choosing (a,f) player 1 believes to obtain at least 15. Since by choosing (a,e) he believes to get at most 5, and by choosing b he believes to get 10, the strategies (a,e) and b are suboptimal for player 1 at h_1. Thus,

$$S_1^{1.3} = \{(a,f)\}.$$

At h_4, player 3 believes that player 2 will choose strategy (d,k,m). But then, the only optimal strategy for player 3 at h_4 is h, and hence

$$S_3^{1.3} = \{h\}.$$

Moreover, $S_2^{1.3} = S_2^{1.2}$.

Round 4. We have that

$$\begin{gathered} B_1^{1.4} = \{b_1 \in B_1^{1.3} \mid b_1 \text{ strongly believes } \{(c,(d,k,m)\} \times \{h\}\} \\ = \{b_1 \in B_1 \mid b_1(h_1)(\{(c,(d,k,m)\} \times \{h\}) = b_1(h_3)(\{(c,(d,k,m)\} \times \{h\}) = 1\} \end{gathered}$$

$$\begin{gathered} B_2^{1.4} = \{b_2 \in B_2^{1.3} \mid b_2 \text{ strongly believes } \{(a,f)\} \times \{h\} \\ = \{b_2 \in B_2 \mid b_2(h_2)(\{(a,f)\} \times \{h\}) = b_2(h_5)(\{(a,f) \times \{h\}) = 1\}, \end{gathered}$$

whereas

$$B_3^{1.4} = \{b_3 \in B_3^{1.3} \mid b_3 \text{ strongly believes } \{(a,f)\} \times \{(c,(d,k,m)\}\} = B_3^{1.3}.$$

Hence, at h_2 player 2 must believe that player 1 chooses (a, f). But then, among the strategies in $S_2^{1.3}$, the only optimal strategy for player 2 at h_2 is (d, k, m). Hence,

$$S_2^{1.4} = \{(d, k, m)\}$$

whereas $S_1^{1.4} = S_1^{1.3} = \{(a, f)\}$ and $S_3^{1.4} = S_3^{1.3} = \{h\}$.

Round 5. We finally have that

$$\begin{aligned} B_1^{1.5} &= \{b_1 \in B_1^{1.4} \mid b_1 \text{ strongly believes } \{(d, k, m)\} \times \{h\}\} \\ &= \{b_1 \in B_1 \mid b_1(h_1)(\{(d, k, m)\} \times \{h\}) = b_1(h_3)(\{(d, k, m)\} \times \{h\}) = 1\} \end{aligned}$$

and

$$\begin{aligned} B_2^{1.5} &= \{b_2 \in B_2^{1.4} \mid b_2 \text{ strongly believes } \{(a, f)\} \times \{h\} \\ &= \{b_2 \in B_2 \mid b_2(h_2)(\{(a, f)\} \times \{h\}) = b_2(h_5)(\{(a, f) \times \{h\}) = b_2(h_6)(\{b\} \times \{(i, n)\}) = 1\}, \end{aligned}$$

whereas

$$\begin{aligned} B_3^{1.5} &= \{b_3 \in B_3^{1.4} \mid b_3 \text{ strongly believes } \{(a, f)\} \times \{(d, k, m)\}\} \\ &= \{b_3 \in B_3 \mid b_3(h_4)(\{b\} \times \{(d, k, m)\} = 1\}. \end{aligned}$$

This is where the procedure terminates. In particular, we see that the unique forward and backward rationalizable strategies for the players are (a, f), (d, k, m) and h, respectively.

4 Epistemic Characterization

In this section we investigate what the concept of forward and backward rationalizability entails in terms of reasoning. To this purpose, we offer epistemic conditions on the players' belief hierarchies such that the optimal strategies under these belief hierarchies are precisely the forward and backward rationalizable strategies. Before doing so, we first recall the definition of a (universal) type space for dynamic games, and subsequently formalize the notion of strong belief and optimal choice for types in a type space.

4.1 Type Space

The epistemic conditions we introduce will impose restrictions on the belief hierarchies that the players may have. Such belief hierarchies may conveniently be encoded by means of *types* in a type space. To formalize a type space, we need the following definition and pieces of notation. A topological space $(X, \mathcal{O})$ is called *Polish* if it is separable and completely metrizable. By $\Sigma(X)$ we

denote the Borel σ-algebra on X, that is, the smallest σ-algebra that contains all open sets, whereas $\Delta(X)$ denotes the set of all probability measures on $(X, \Sigma(X))$. We endow $\Delta(X)$ with the smallest topology $\mathcal{O}_{\Delta(X)}$ such that each of the sets $\{\mu \in \Delta(X) | \int_X f d\mu \in O\}$ is open in $\Delta(X)$, where f runs over all bounded continuous functions $f : X \to \mathbf{R}$ and O runs over all open subsets of the reals. By Kechris (1995), Theorem 17.23, $(\Delta(X), \mathcal{O}_{\Delta(X)})$ is again a Polish space. We then consider $\Delta(X)$ as a measurable space that is endowed with the Borel σ-algebra (generated by $\mathcal{O}_{\Delta(X)}$). It is a well-known fact that a continuous map between two topological spaces is measurable if both of these spaces are endowed with their respective Borel σ-algebras.

Definition 4.1 (Type space) *A **type space** $\mathcal{T} = ((T_i, \mathcal{O}_i), \beta_i)_{i \in I}$ specifies, for every player i,*

(a) a Polish type space $(T_i, \mathcal{O}_i)$, and

(b) a continuous belief mapping β_i, which assigns to every type $t_i \in T_i$ and information set $h \in H_i$ a probabilistic belief $\beta_i(t_i, h) \in \Delta(S_{-i}(h) \times T_{-i})$.

Moreover, the types must satisfy Bayesian updating whenever possible, that is, for every player i, every type $t_i \in T_i$, and every two information sets $h, h' \in H_i$ where h' follows h and $\beta_i(t_i, h)(S_{-i}(h') \times T_{-i}) > 0$, we have that

$$\beta_i(t_i, h')(\{s_{-i}\} \times E_{-i}) = \frac{\beta_i(t_i, h)(\{s_{-i}\} \times E_{-i})}{\beta_i(t_i, h)(S_{-i}(h') \times T_{-i})}$$

for every $s_{-i} \in S_{-i}(h')$ and every $E_{-i} \in \Sigma(T_{-i})$.

For our epistemic characterization we need to work with a *universal* type space. To explain what it is, we must first introduce the notion of a *type morphism.*

Definition 4.2 (Type morphism) *Consider two type spaces $\mathcal{T} = ((T_i, O_i), \beta_i)_{i \in I}$ and $\mathcal{T}' = ((T_i', O_i'), \beta_i')_{i \in I}$. A **type morphism** from $\mathcal{T}$ to $\mathcal{T}'$ is a tuple $(f_i)_{i \in I}$ of continuous functions $f_i : T_i \to T_i'$ such that, for every player i, every type $t_i \in T_i$, and every information set $h \in H_i$ we have that*

$$\beta_i'(f_i(t_i), h)(\times_{j \neq i}(\{s_j\} \times E_j')) = \beta_i(t_i, h)(\times_{j \neq i}(\{s_j\} \times f_j^{-1}(E_j'))$$

for every opponents' strategy combination $(s_j)_{j \neq i} \in S_{-i}(h)$ and every measurable set $\times_{j \neq i} E_j' \subseteq \times_{j \neq i} T_j'$ of opponents' type combinations.

A type space is then called *universal* if every other type space can be uniquely embedded into it by means of a type morphism.

Definition 4.3 (Universal type space) *A type space $\mathcal{T}$ is **universal**[7] if for every other type space $\mathcal{T}'$ there is a unique type morphism from $\mathcal{T}'$ to $\mathcal{T}$.*

It turns out that every two universal type spaces are isomorphic. As such, we can speak about *the* universal type space. Battigalli and Siniscalchi (1999), Guarino (2022) and Fukuda (2023) have shown that we can always construct a universal type space in our setting.

[7] In the literature, such type spaces are sometimes called *terminal.*

4.2 Strong Belief

Consider a type space $\mathcal{T} = ((T_i, \mathcal{O}_i), \beta_i)_{i \in I}$, a type $t_i \in T_i$ and an event $E_{-i} \in \Sigma(S_{-i} \times T_{-i})$. The type t_i is said to *strongly believe* the event E_{-i} if it assigns probability 1 to the event whenever possible. That is,

$$\beta_i(t_i, h)(E_{-i}) = 1 \text{ at all } h \in H_i \text{ where } E_{-i} \cap (S_{-i}(h) \times T_{-i}) \neq \emptyset.$$

4.3 Optimal Choice

Consider a type $t_i \in T_i$, a strategy $s_i \in S_i$ and an information set $h \in H_i(s_i)$ that can possibly be reached by s_i. Then, we denote by

$$u_i(s_i, t_i, h) := \sum_{s_{-i} \in S_{-i}(h)} \beta_i(t_i, h)(\{s_{-i}\} \times T_{-i}) \cdot u_i(z(s_i, s_{-i}))$$

the *expected utility* induced by s_i at h for the type t_i. The strategy s_i is *optimal* for the type t_i at information set h if $u_i(s_i, t_i, h) \geq u_i(s_i', t_i, h)$ for all other strategies $s_i' \in S_i(h)$. For a given period m, we say that strategy s_i is optimal for the type t_i *from period m onwards* if for every period $\tau \geq m$, and every information set $h \in H_i(s_i) \cap H^\tau$, the strategy s_i is optimal for type t_i at h.[8] For a given set of types $\hat{T}_i \in \Sigma(T_i)$, we denote by

$$(S_i \times \hat{T}_i)^{rat,m} := \{(s_i, t_i) \in S_i \times \hat{T}_i \mid s_i \text{ is optimal for } t_i \text{ from period } m \text{ onwards}\}$$

the event that player i chooses rationally from period m onwards and that i's type belongs to $\hat{T}_i$.

The following result states that the event of choosing rationally from a certain period onwards is always a "well-behaved" set.

Lemma 4.1 (Rationality is a measurable event) *Suppose that $\hat{T}_i$ is a closed (measurable) subset of T_i. Then, the set $(S_i \times \hat{T}_i)^{rat,m}$ is a closed (measurable) subset of $S_i \times T_i$.*

This result will be important for guaranteeing that the epistemic conditions below are all well-defined. It will also play a key role in the proof of our epistemic characterization.

4.4 Epistemic Characterization

The epistemic conditions we impose on the players' types are as follows: First, we focus on the last period M where players have to move. A player must $(M.1)$ strongly believe in the event that every opponent chooses rationally from period M onwards, $(M.2)$ strongly believe in the event that every opponent chooses rationally from period M onwards and that every opponent satisfies $(M.1)$, and

[8] Note that if $h \notin H_i(s_i) \cap H^\tau$ for all $\tau \geq m$, then s_i is (vacuously) optimal for every type of player i from period m onwards.

so on. These conditions together yield *common backward strong belief in rationality from period M onwards.* We refer to this event as (M). In fact, since every player moves at most once at period M, event (M) is equivalent to *common belief in rationality at period M.*

We then move to period $M-1$. A player must $(M-1.1)$ strongly believe in the event that every opponent chooses rationally from period $M-1$ onwards and that every opponent satisfies (M). Moreover, a player must $(M-1.2)$ strongly believe in the event that every opponent chooses rationally from period $M-1$ onwards and that every opponent satisfies $(M-1.1)$, and so on. These conditions together yield *common backward strong belief in rationality from period $M-1$ onwards.*

We then continue in this fashion until we reach the beginning of the game. The final restrictions on the types are called *common backward strong belief in rationality.*

Definition 4.4 (Common backward strong belief in rationality) *For every period m, number $k \in \{0,1,...\}$ and player i, we define the sets of types $T_i^{m.k}$ that express k-fold backward strong belief in rationality from period m onwards. These sets $T_i^{m.k}$ are inductively defined as follows.*

***Period** M. Set $T_i^{M.0} := T_i$ for every player i. For every $k \geq 1$, inductively define*

$$T_i^{M.k} := \{t_i \in T_i^{M.k-1} \mid t_i \text{ strongly believes } \times_{j \neq i} (S_j \times T_j^{M.k-1})^{rat,M}\}.$$

Set $T_i^M := \cap_{k \geq 0} T_i^{M.k}$ for every player i.

***Period** $m \leq M-1$. Set $T_i^{m.0} := T_i^{m+1}$ for every player i. For every $k \geq 1$, inductively define*

$$T_i^{m.k} := \{t_i \in T_i^{m.k-1} \mid t_i \text{ strongly believes } \times_{j \neq i} (S_j \times T_j^{m.k-1})^{rat,m}\}.$$

Set $T_i^m := \cap_{k \geq 0} T_i^{m.k}$ for every player i.

For a given period m and round k, a type t_i is said to express up to k-fold backward strong belief in rationality from period m onwards if $t_i \in T_i^{m.k}$. The type t_i is said to express common backward strong belief in rationality from period m onwards if $t_i \in T_i^m$. The type t_i is said to express common backward strong belief in rationality if $t_i \in T_i^L$, where L is the first period in the game.

The following result guarantees that the epistemic conditions imposed above lead to "well-behaved" sets.

Lemma 4.2 (Epistemic conditions lead to closed sets) *Each of the sets $T_i^{m.k}$ and T_i^m above is a closed subset of T_i.*

Let us now have a closer look at the epistemic conditions above. The conditions imply that at every information set where a player has to move, he looks for the earliest period m and the highest degree k such that it is possible to believe that (i) every player chooses rationally from period m onwards and expresses common backward strong belief in rationality from period m onwards, and

(ii) every player chooses rationally from period $m-1$ onwards and expresses up to k-fold backward strong belief in rationality from period $m-1$ onwards. Moreover, he *will* then believe (i) and (ii). This may be viewed as a *best rationalization principle* for the epistemic concept above.

From this best rationalization principle it is clear that epistemic priority is given to backward induction reasoning: If a player is at an information set, he first looks for the earliest period m such that it is possible to believe that every player chooses rationally from period m onwards and expresses common backward strong belief in rationality from period m onwards. In that case, the player *will* express common backward strong belief in rationality from period m onwards, and hence will believe, in particular, that every opponent will choose rationally from period m onwards. Only afterwards will he think about period $m-1$, and look for the highest degree k such that it is possible to believe that, in addition, every player chooses rationally from period $m-1$ onwards and expresses up to k-fold backward strong belief in rationality from period $m-1$ onwards.

The following result shows that the epistemic conditions in *common backward strong belief in rationality* single out precisely those strategies that are *forward and backward rationalizable.*

Theorem 4.1 (Epistemic characterization) *Consider the universal type space $\mathcal{T} = ((T_i, \mathcal{O}_i), \beta_i)_{i \in I}$. Then, for every player i and strategy $s_i \in S_i$, the following holds:*

(a) strategy s_i is forward and backward rationalizable, if and only if, s_i is optimal from the first period onwards for a type $t_i \in T_i$ that expresses common backward strong belief in rationality,

(b) if $m \leq M-1$ then $s_i \in S_i^{m.0}$, if and only if, s_i is optimal from period $m+1$ onwards for a type $t_i \in T_i^{m+1}$ that expresses common backward strong belief in rationality from period $m+1$ onwards, and

(c) if $k \geq 0$ then $s_i \in S_i^{m.k+1}$, if and only if, s_i is optimal from period m onwards for a type $t_i \in T_i^{m.k}$ that expresses up to k-fold backward strong belief in rationality from period m onwards.

In particular, since we know from Theorem 3.1 that forward and backward rationalizable strategies always exist, it follows that there is always a type that expresses common backward strong belief in rationality. That is, the system of epistemic conditions we offer never leads to logical contradictions.

A major difference with strong rationalizability is that forward and backward rationalizability requires players to do forward induction reasoning from a certain period onwards, in a backward inductive fashion. Strong rationalizability, in contrast, always requires players to do the forward induction reasoning in the whole game, that is, from the first period onwards.

As such, we can also consider a *bounded rationality* version of forward and backward rationalizability in which players only do the forward induction reasoning from period M onwards, from period $M-1$ onwards, until we reach period m. Players would thus not actively reason about choices that are made before period m. Parts (b) and (c) in Theorem 4.1 reveal what has to be imposed, in terms of reasoning, to establish such a bounded rationality variant.

5 Relation with Backwards Rationalizability

In this section we start by showing that our concept of forward and backward rationalizability is a refinement of backwards rationalizability in terms of strategies. In epistemic terms, this means that our concept will always reason within the bounds of *common belief in future rationality* (Perea (2014)). It is shown that the reasoning of strong rationalizability may be in conflict with common belief in future rationality, and we explain why in some situations this may lead to unreasonable behavior after observing surprising past moves. We conclude by looking at two alternative procedures that also combine backward and forward induction reasoning, and show by means of examples that these may be different from our procedure in terms of strategies.

5.1 Refinement of Backwards Rationalizability

In the game of Figure 2 we saw that forward and backward rationalizability selects a different strategy for player 2 than strong rationalizability. The reason was that according to the former concept, player 2, at a given information set h, only interprets player 1's past move as a rational move if this is compatible with the completed reasoning from h onwards. This shows that forward and backward rationalizability is, above all, a forward looking concept, and thus gives priority to backward induction reasoning.

This intuition will be confirmed in this section, where we show that forward and backward rationalizability always yields a refinement – both in terms of strategies and beliefs – of the *backwards rationalizability* procedure, as defined in Penta (2015) and Perea (2014).

We are now ready to state the announced result.

Theorem 5.1 (Relation with backwards rationalizability) *Every strategy and conditional belief vector that is forward and backward rationalizable, is also backwards rationalizable.*

That is, our concept of forward and backward rationalizability will always reason in line with backwards rationalizability, even if a player is surprised by some opponents' past moves. However, as the game in Figure 2 shows, our concept may be more restrictive since it additionally imposes some forward induction reasoning. Indeed, in that game backwards rationalizability allows player 2 to choose either d or (c, h), as it does not perform any forward induction reasoning. Our concept, in turn, uniquely selects player 2's strategy (c, h).

5.2 Belief in Future Rationality

In Perea (2014) it is shown that backwards rationalizability can be epistemically characterized by the conditions of *common belief in future rationality,* stating that a player always believes that his opponents will choose rationally now and in the future, that a player always believes that his

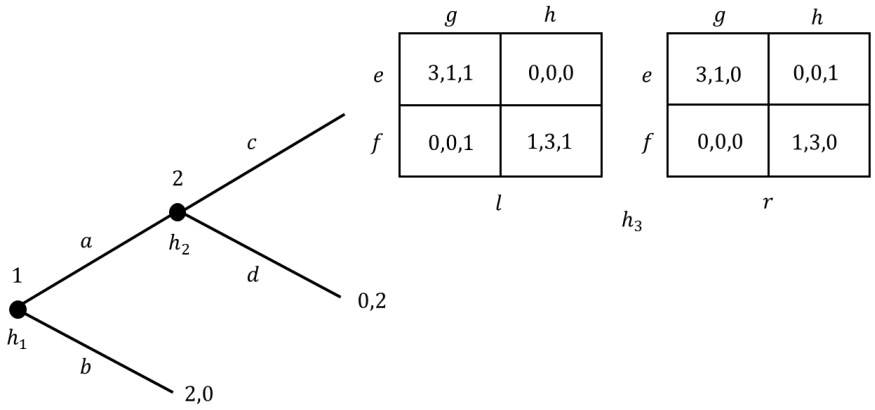

Figure 4: Double outside option game with a bet for the third player

opponents always believe that the other players will chooses rationally now and in the future, and so on. Hence, even if a player is suprised by an opponent's past move, which may possibly lead him to conclude that his opponent has chosen irrationally in the past, he will still believe that the same opponent will "recover" and choose rationally from now on.

As, by Theorem 5.1, forward and backward rationalizability refines the reasoning of backwards rationalizability, it follows that the former concept always reasons within the bounds set by common belief in future rationality. In particular, a player will only interpret an opponent's past move as a signal for the opponent's future behavior – a typical forward inductive inference – if this is in accordance with common belief in future rationality. If, for instance, the opponent's observed past move could only be rational if the opponent believes that some other players will choose irrationally in the future, then, according to our concept, the player should not view this move as a credible signal for future behavior.

In this sense, our concept is fundamentally different from strong rationalizability. To further illustrate this, consider the game in Figure 4. This is a three-player double outside game, where the third player has to "bet" on the behavior of the first two players after they have both foregone the outside option. In that case, player 1 must choose between e and f, player 2 must choose between g and h, whereas player 3 must choose between l and r (left matrix or right matrix). These choices by players 1, 2 and 3 are assumed to occur simultaneously. Note that r will only be optimal for player 3 if he believes, with sufficiently high probability, that players 1 and 2 miscoordinate on e and h at information set h_3.

According to strong rationalizability, player 3 will conclude at h_3 that players 1 and 2 will go for e and h, respectively, and therefore player 3 will choose r. To see this, note that a for player 1 can only be optimal at h_1 if he chooses e at h_3, and that c for player 2 can only be optimal at h_2 if he

chooses h at h_3.

But assume now that player 1 would choose optimally at h_3 given his beliefs there. According to strong rationalizability, player 1 will believe at h_3 that player 2 chooses h, for the same reasons as outlined above. If player 1 chooses optimally at h_3, then he would go for f and not e, rendering r suboptimal for player 3. In that sense, the choice r selected by strong rationalizability is a risky choice for player 3, as this bet is based on the assumption that player 1 will not choose optimally given his belief at h_3. That is, player 3 would not believe in player 1's present rationality at h_3, and his reasoning would thus contradict common belief in future rationality.

The concept of forward and backward rationalizability prescribes a completely different line of reasoning for player 3 here. If player 3 finds himself at h_3, then he first asks whether there is a plausible theory from information set h_2 onwards that could explain the event of reaching h_3. Such a theory can indeed be found: For player 2 it is only optimal to choose c at h_2 if he would choose h at h_3. Therefore, player 3 will believe at h_3 that player 2 will choose h, and will believe that player 1 also believes at h_3 that player 2 chooses h. Assuming that player 1 chooses optimally at h_3, player 3 then believes that player 1 chooses f at h. Hence, by analyzing the game from h_2 onwards, player 3 comes up with a unique prediction for the behavior of players 1 and 2 at h_3, which is that they choose f and h, respectively.

Clearly, this prediction cannot be refined any further by subsequently analyzing the game from h_1 onwards, and therefore forward and backward rationalizability leads player 3 to believe that players 1 and 2 will choose f and h at h_3, and player 3 will thus choose l. Note that player 1's choice f is optimal at h_3 if he believes that player 2 chooses h there, and *vice versa.* As such, player 3's reasoning about the behavior of players 1 and 2 is in line with common belief in future rationality, contrary to player 3's reasoning prescribed by strong rationalizability.

5.3 Epistemic Priority

One could argue that in the concept we propose, we give epistemic priority to backward induction reasoning over forward induction reasoning. This may be seen, for instance, from Theorem 5.1 which shows that our concept, in terms of strategies and beliefs, is a refinement of the backward induction concept of backwards rationalizability, but not of strong rationalizability.

It may also be seen epistemically, by looking at Theorem 4.1: We first impose common strong belief in rationality from the last period onwards, and keep these restrictions when we move to restrictions on reasoning from the penultimate period onwards. In turn, the restrictions on the reasoning from the penultimate period onwards are maintained when restricting the reasoning from earlier periods onwards, and so on. As such, a player will always believe, at every period, that his opponents will choose rationally in the periods that lie ahead – a typical backward induction restriction.

But we could also change the epistemic priority, by first applying the strong rationalizability procedure, and subsequently refining it by the backwards rationalizability procedure. In that alter-

native procedure we would thus give epistemic priority to forward induction reasoning over backward induction reasoning.

This concept would be different from ours, at least in terms of strategies and beliefs. Consider, for instance, the game from Figure 1, where our concept would yield the strategy c for player 2, whereas the alternative procedure would uniquely select strategy (d, g) for player 2. To see this, note that backwards rationalizability always leads to the unique backward induction strategies in perfect information games without relevant ties, like the one in Figure 1. As player 2's backward induction strategy is c, and our concept is a refinement of backwards rationalizability in terms of strategies, our concept will uniquely select c as well. On the other hand, we have seen in the introduction that player 2's unique strongly rationalizable strategy is (d, g). As the alternative procedure is a refinement of strong rationalizability in terms of strategies, it will uniquely select (d, g) also.

For a concept that combines forward and backward induction reasoning, one could also proceed alternatively, by first applying the backwards rationalizability procedure to the whole game, until we can go no further, after which it is refined by the steps in the strong rationalizability procedure. This would correspond to an instance of Δ-*rationalizability* (Battigalli (2003), Battigalli and Siniscalchi (2003)) where Δ consists of the restrictions on beliefs imposed by backwards rationalizability. Like with our procedure, this would also correspond to a scenario where epistemic priority is given to backward induction reasoning, but in a more extreme fashion than we do. Indeed, in the alternative procedure we would first exhaust all the backward induction reasoning in the whole game, after which we exclusively turn to forward induction reasoning in the whole game.

For reasons that will become clear in the following section, the alternative procedure will be equivalent to ours in terms of outcomes. Moreover, like our procedure, it will also refine backwards rationalizability in terms of strategies. However, as the example in Figure 5 will show, both concepts can be different in terms of strategies.

Consider the game from Figure 5. In the alternative procedure, we would start by applying the backwards rationalizability procedure to the whole game. We proceed in a backward inductive fashion here, by first considering the last information set h_4, where nothing can be eliminated. At h_3, we eliminate (In, f) for player 2, after which we can eliminate (In, c) for player 1 at h_2. Finally, we eliminate (In, r) for player 3 at h_1. The backwards rationalizable strategies are thus Out, (In, a) and (In, b) for player 1, Out, (In, d) and (In, e) for player 2, and Out and (In, l) for player 3.

If we take this as an input for the strong rationalizability procedure, then in round 1 of the strong rationalizability procedure we can eliminate (In, b) for player 1 and (In, d) for player 2. Indeed, at h_2 player 1 must believe that player 2 chooses Out, (In, d) or (In, e) and that player 3 chooses (In, l). Hence, player 1 expects at most 1 by choosing (In, b) there. Also, player 2 must believe at h_3 that player 1 will choose (In, a) or (In, b) and that player 3 will choose (In, l). As such, player 2 expects at most 1 by choosing (In, d) there.

In round 2 we can then eliminate, for similar reasons, (In, a) for player 1 and (In, e) for player 2. In round 3 we can finally eliminate (In, l) for player 3. Indeed, player 3 must believe at h_1 that player 1 chooses Out, which yields Out as the only optimal strategy for player 3 at h_1.

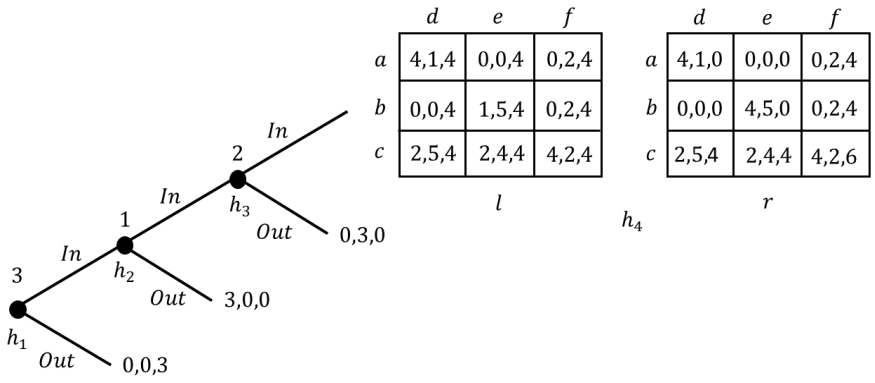

Figure 5: Triple outside option game

The alternative procedure, where we first apply the backwards rationalizability procedure and then strong rationalizability, thus yields the strategy Out for player 1, Out for player 2, and Out for player 3.

Let us now apply our procedure to this game. Applying the strong rationalizability procedure to the subgame starting at h_4 yields no eliminations. If we start at h_3, then we can only eliminate the strategy (In, f) for player 2. If we start at h_2, then in round 1 we eliminate the strategy (In, c) for player 1, as he expects to get at most 2 by choosing (In, c) at h_2. In round 2 we would then eliminate the strategy (In, d) for player 2 and the strategy (In, r) for player 3. Indeed, player 2 expects to get at most 1 by playing (In, d), whereas for player 3 choosing l has become better than choosing r at h_4. In round 3, we would then eliminate (In, a) and (In, b) for player 1, since he expects to obtain no more than 1 by choosing either of these two strategies. Finally, we would move to the game starting at h_1, where we can eliminate the strategy (In, l) for player 3. Our concept would thus yield the strategy Out for player 1, the strategies Out and (In, e) for player 2, and the strategy Out for player 3. As the strategy (In, e) for player 2 was not selected by the alternative procedure, both concepts differ in terms of strategies.

The reason for why our concept allows for player 2's strategy (In, e) but the alternative procedure does not, is the following: The alternative procedure starts by eliminating the strategies (In, f) for player 2, (In, c) for player 1 and (In, r) for player 3. It would then proceed by applying strong rationalizability to the whole game, so that player 2 will conclude at h_3 that player 1 must be choosing (In, a). As a consequence player 2 must choose Out at h_3 according to the alternative procedure.

Our concept proceeds differently: It also starts by eliminating (In, f) for player 2 and (In, c) for player 1. But then, by reasoning from h_2 onwards, we would eliminate (In, d) for player 2 and

(In, r) for player 3. If player 1 believes at h_2 that player 2 will no longer choose (In, d) and (In, f), both (In, a) and (In, b) become suboptimal for player 1 at h_1. As (In, c) was already eliminated for player 1 at an earlier stage, player 2 is no longer forced to discriminate between (In, a) and (In, b), and hence player 2 may still believe at h_3 that player 1 chooses (In, b). Hence, player 2 may still rationally choose (In, e) at h_3 according to our concept.

In this example, we thus see that the alternative concept is more restrictive, in terms of strategies, than ours. The reverse may also be true, as the example from Figure 2 shows. We have already seen in the introduction and Section 3.4 that our concept uniquely selects the strategies b for player 1 and (c, h) for player 2.

Suppose now that we would run the alternative procedure. By applying backwards rationalizability first, we would start by eliminating strategy (c, g) for player 2 at h_2, after which we would eliminate (a, e) and (a, f) for player 1 at h_1. Indeed, if player 1 believes at h_1 that player 2 will not choose (c, g), then choosing a can give him at most 1. Hence, the backwards rationalizable strategies are b for player 1, and (c, h) and d for player 2. If we use this as the input for the strong rationalizability procedure, then at h_2 player 2 can no longer discriminate between (a, e) and (a, f) for player 1, and hence player 2 may believe at h_2 that player 1 chooses (a, e) or (a, f). As such, both d and (c, h) can be optimal for player 2 at h_2, which means that the alternative concept would select both d and (c, h) for player 2.

The reason for this difference is similar to above, but now with the roles of the two concepts reversed: Under our concept, player 2 will certainly believe at h_2 that player 1 chooses (a, f) and not (a, e), and therefore only (c, h) is optimal for player 2. Under the alternative concept, player 2 is no longer forced to discriminate between (a, e) and (a, f), which leaves both d and (c, h) as valid options for player 2.

The last alternative procedure described above thus gives epistemic priority to backward induction reasoning over forward induction reasoning. Similar approaches have been adopted in the equilibrium refinements literature from the eighties and early nineties, where the backward induction concept of sequential equilibrium has been refined by imposing forward induction restrictions. See, for instance, *justifiable sequential equilibrium* (McLennan (1985)), *forward induction equilibrium* (Cho (1987)) and *stable sets of beliefs* (Hillas (1994)).

The issue of *epistemic priority* is explored in depth by Catonini (2019), who proposes the concept of *selective rationalizability.* It proceeds by first applying strong rationalizability, after which it is refined by imposing (common strong belief in) some exogenously given restrictions on beliefs.[9] However, it could happen that these exogenous restrictions are incompatible with the restrictions imposed by strong rationalizability. This will typically be the case when the exogenous restrictions are taken to be the restrictions imposed by backwards rationalizability, because in many dynamic games these restrictions go against the restrictions of strong rationalizability.

[9]Instead of starting with strong rationalizability one could also start with some other concept, and then impose exogenous restrictions on the beliefs. Catonini (2019) uses strong rationalizability as the focal concept here, but his analysis allows it to be replaced by any other rationalizability concept for dynamic games as well.

The first alternative procedure described above, where we started with the strong rationalizability procedure, and subsequently refined it with the backwards rationalizability procedure, is, strictly speaking, not a selective rationalizability procedure in the sense of Catonini (2019): The restrictions of backwards rationalizability are not being imposed as *exogeneous* restrictions after completing the strong rationalizability procedure, but rather as "optional" restrictions meant to refine, *whenever possible*, the reasoning of strong rationalizability.

6 Relation with Strong Rationalizability

In this section we show that the forward and backward rationalizability procedure is equivalent, in terms of outcomes, to the strong rationalizability procedure. To prove this result we use the *strong belief reduction operator* from Perea (2018), transforming every product of strategy sets into a product of smaller strategy sets. Perea (2018) showed, for the class of dynamic games with *observed past choices*, that (a) strong rationalizability can be characterized by the iterated application of the strong belief reduction operator, and (b) the strong belief reduction operator is order independent with respect to outcomes, that is, the final set of induced outcomes will not change if during the iterated application of the strong belief operator we opt for a "slower" elimination order in which we do not always eliminate all strategies we can.

In this section we argue that the results (a) and (b) carry over to the more general class of dynamic games we consider in this paper, which allows for *unobserved past choices*. Moreover, we show that (c) the forward and backward rationalizability procedure corresponds to a specific "slower" elimination order of the iterated application of the strong belief reduction operator. By combining the properties (a), (b) and (c) it then follows that the strong rationalizability procedure and the forward and backward rationalizability procedure always induce the same set of outcomes.

6.1 Strong Belief Reduction Operator

A *product of strategy sets* is a set $D = \times_{i \in I} D_i$ where $D_i \subseteq S_i$ for every player i. A *reduction operator* r assigns to every product of strategy sets D a subset $r(D) \subseteq D$, where $r(D)$ is again a product of strategy sets. In the sequel, we always assume that we are talking about products of strategy sets. A set E with $r(D) \subseteq E \subseteq D$ is a *partial reduction* of D. For a given $k \in \mathbf{N}$, let $r^k(D)$ be the k-fold application of the operator r to D.

For a given product of strategy sets D, let $H(D)$ be the collection of information sets that are reached by some strategy profile in D.

Definition 6.1 (Strong belief reduction operator) *The* ***strong belief reduction operator***

sb assigns to every product of strategy sets $D = \times_{i \in I} D_i$ *the set* $sb(D) = \times_{i \in I} sb_i(D)$, *where*

$$sb_i(D) := \{s_i \in D_i \mid \text{there is some } b_i \in B_i \text{ that strongly believes } D_{-i} \text{ such that } s_i \text{ is optimal for } b_i \text{ at every } h \in H(D) \cap H_i(s_i)\}.$$

Recall that $H_i(s_i)$ is the collection of information sets for player i that can be reached by s_i. In Perea (2018), Theorem 2.1, it is shown that for dynamic games with observed past choices, the strongly rationalizable strategies are obtained by iteratedly applying the strong belief reduction operator to the full set of strategies. As the proof does not rely on the property of observed past choices, this result holds for all games in our setting as well.

Theorem 6.1 (Characterization of strong rationalizability) *For every* $k \in \mathbf{N}$ *and every player* i, *let* $S_i^{sr,k}$ *be the set of strategies for player* i *that survive round* k *of the strong rationalizability procedure. Let* $S^{sr,k} = \times_{i \in I} S_i^{sr,k}$ *and* $S = \times_{i \in I} S_i$. *Then,* $S^{sr,k} = sb^k(S)$ *for every* $k \in \mathbf{N}$.

6.2 Order Independence with Respect to Outcomes

Consider a reduction operator r. An *elimination order* for r is a finite sequence $(D^0, D^1, ..., D^K)$ of products of strategy sets such that (a) $D^0 = S$, (b) $r(D^k) \subseteq D^{k+1} \subseteq D^k$ for every $k \in \{0, ..., K-1\}$, and (c) $r(D^K) = D^K$.

For a product of strategy sets D, let $Z(D)$ be the collection of terminal histories reached by strategy profiles in D.

Definition 6.2 (Order independence with respect to outcomes) *A reduction operator* r *is* ***order independent with respect to outcomes*** *if for every two elimination orders* $(D^0, D^1, ..., D^K)$ *and* $(E^0, E^1, ..., E^L)$ *we have that* $Z(D^K) = Z(E^L)$.

Corollary 3.1 in Perea (2018) states that for all dynamic games with observed past choices, the strong belief reduction operator is order independent with respect to outcomes. As it turns out, the proof in Perea (2018) does not rely on the property of observed past choices, and holds for our class of dynamic games as well.

Theorem 6.2 (Order independence with respect to outcomes) *The strong belief reduction operator sb is order independent with respect to outcomes.*

6.3 Outcome Equivalence with Strong Rationalizability

The following result states that the reduction steps in the forward and backward rationalizability procedure correspond to a specific elimination order of the strong belief reduction operator.

Lemma 6.1 (Procedure as elimination order) *Let $(D^0, D^1, ..., D^K)$ be the products of strategy sets generated by the forward and backward rationalizability procedure in every round.*[10] *Then, $(D^0, D^1, ..., D^K)$ is an elimination order of the strong belief reduction operator.*

By combining Theorem 6.1, Lemma 6.1 and Theorem 6.2, we conclude that strong rationalizability leads to the same set of outcomes as forward and backward rationalizability.

Theorem 6.3 (Outcome equivalence with strong rationalizability) *Let S^{sr} and S^{fbr} be the products of strategy sets induced by the strong rationalizability procedure, and the forward and backward rationalizability procedure, respectively. Then, $Z(S^{sr}) = Z(S^{fbr})$.*

That is, if one would be only interested in the induced outcomes, it makes no difference whether strong rationalizability is used, or forward and backward rationalizability. However, as we have argued before, we believe that the latter concept provides a more compelling theory for how players react to surprises at information sets to which the players initially assign probability zero.

Finally, consider one of the alternative procedures we discussed in Section 5.3, where we first apply the backwards rationalizability procedure, followed by the strong rationalizability procedure. By a proof similar to that of Lemma 6.1, it can be shown that this procedure also corresponds to a specific elimination order of the strong belief reduction operator. But then, it follows in the same way as above that also this procedure is equivalent to strong rationalizability in terms of outcomes. However, recall from Section 5.3 that the two procedures may be different in terms of strategies.

For the other alternative procedure in Section 5.3, where we first apply the strong rationalizability procedure, followed by the backwards rationalizability procedure, things are more complicated, as it does not correspond to a specific elimination order of the strong belief reduction operator. The reason is that after applying the strong rationalizability procedure, which corresponds to recursively applying the strong belief reduction operator at "full speed", the strong belief reduction operator is not able to induce any further eliminations. At this moment it is not clear to us whether both concepts can be different in terms of outcomes. But we have seen in Section 5.3 that both concepts can be different in terms of strategies.

7 Generalization of Battigalli's Theorem

Battigalli (1997) has shown that in every dynamic game with perfect information but without relevant ties, strong rationalizability leads to the unique backward induction outcome. Alternative proofs can be found in Catonini (2020), Chen and Micali (2013), Heifetz and Perea (2015) and Perea (2018).

Catonini (2020) and Perea (2017) provide generalizations of this result, by showing that in every dynamic game with *observable past choices* (but allowing for simultaneous moves), strong rationalizability refines, in terms of outcomes, the concepts of backwards rationalizability and backward dominance, respectively.

However, one of the most attractive properties of strong rationalizability is that in games with *imperfect* information (that is, where some past choices are unobservable), it allows an active player at an information set to exclude some of the nodes of this information set from consideration. This is so, since reaching those nodes would imply a lower degree of rationality for some players that moved in the past than the remaining "more rational" nodes. Therefore it is of importance to understand the relationship between forward and backward induction especially for games with *imperfect* information. As is well-known, there are games that allow for more backwards rationalizable outcomes than extensive-form outcomes.[11] To the best of our knowledge, it was – up to now – unknown whether for games with *imperfect* information strongly rationalizable outcomes are always backwards rationalizable outcomes.

It turns out that in every dynamic game considered in our setup, that is, also in games with imperfect information, every strongly rationalizable outcome is indeed backwards rationalizable. This follows immediately from our Theorems 5.1 and 6.3: Take an outcome induced by strong rationalizability. Then, by Theorem 6.3, this outcome will also be induced by forward and backward rationalizability. As, by Theorem 5.1, forward and backward rationalizability refines backwards rationalizability in terms of strategies, it follows that this outcome is also induced by backwards rationalizability. We thus obtain the following result.

Corollary 7.1 (Generalization of Battigalli's theorem) *Let S^{sr} and S^{br} denote the products of strategy sets induced by the strong rationalizability procedure and the backwards rationalizability procedure, respectively. Then, $Z(S^{sr}) \subseteq Z(S^{br})$.*

This result, in turn, implies Battigalli's theorem, as in every dynamic game with perfect information but without relevant ties, the concept of backwards rationalizability leads to the unique backward induction strategies, and thus, in particular, to the unique backward induction outcome. The existence of strongly rationalizable outcomes now implies that this must also be the unique strongly rationalizable outcome.

Corollary 7.1 could also be established by using tools from Chen and Micali (2013)[12]: It can be shown that backwards rationalizability corresponds to a possible, non-finished elimination order of the iterated conditional dominance procedure by Shimoji and Watson (1998). Since Shimoji and Watson (1998) prove that the latter procedure is equivalent to strong rationalizability, and Chen and Micali (2013) show that this procedure is order independent with respect to outcomes, the statement from the corollary follows.

[11] See, for instance, the classical Battle-of-the-Sexes game with an outside option, which is the game that starts at h_2 in Figure 2. In that game, extensive-form rationalizability uniquely yields the forward induction outcome $(c, (f, h))$, whereas backwards rationalizability also allows for the outcomes $(c, (e, h))$ and d.

[12] We thank Pierpaolo Battigalli and Emiliano Catonini for pointing this out to us.

8 Supergame Monotonicity

In this section we introduce the principle of *supergame monotonicity*. Intuitively, it states that, whenever a player is informed that the game was actually preceded by some earlier moves he was initially unaware of, then this new piece of information should only lead him to possibly *refine* his reasoning, but never to *completely overthrow* the reasoning he applied before. We start by illustrating the idea by means of an example, after which we formally state the principle. It is then shown that our concept of forward and backward rationalizability satisfies this principle, whereas strong rationalizability violates it. We conclude by showing that backwards rationalizability satisfies supergame monotonicity.

From these findings it follows that, upon learning that the game Γ was preceded by some earlier moves, and thus that the actual game is $\hat{\Gamma}$ and not Γ, forward and backward rationalizability will (i) refine the outcomes induced by strong rationalizability in the smaller game Γ, and (ii) refine the strategies induced by backwards rationalizability in the smaller game Γ. In that sense, the forward and backward rationalizability concept will always be in line with forward induction reasoning and backward induction reasoning, even if new information about past moves comes in.

8.1 Example

Consider a game Γ, and suppose that the players in Γ receive some additional information about the history that led to the game Γ. Formally, Γ will be embedded into a larger game $\hat{\Gamma}$, which includes Γ as a subgame. We call $\hat{\Gamma}$ a *supergame* of Γ.

As an illustration, consider Figure 1 and the game Γ that starts at information set h_2. Now suppose that player 2 receives additional information about the history that led to this game. More specifically, player 2 learns that player 1 could have chosen an outside option, a, which would have given him a utility of 5, but player 1 declined this option. Formally, player 2 learns that the game is expanded to the supergame $\hat{\Gamma}$ which starts at information set h_1.

Supergame monotonicity then states that this additional information may possibly *refine* the reasoning of the players, but will *never overrule* it. More precisely, every strategy that can rationally be chosen in the smaller game Γ with the reasoning induced by the *new* information should also be allowed under the *original* reasoning, *without* this new information.

As it turns out, the concept of *strong rationalizability* violates supergame monotonicity with respect to the games Γ and $\hat{\Gamma}$ in Figure 1. To see this, consider first the game Γ that starts at h_2. According to strong rationalizability, player 1, at h_3, must believe that player 2's action d is part of an optimal strategy. This is only possible if player 2 would choose g at h_4. Thus, player 1 must believe at h_3 that player 2 chooses g at h_4, and therefore player 1 must choose e at h_3. Player 2, anticipating this reasoning by player 1, will believe at h_2 that player 1 chooses e at h_3.

Assume now that player 2 is informed about the fact that player 1 did not choose the outside option a before h_2 is reached. That is, player 2 learns that the actual game is $\hat{\Gamma}$, starting at h_1. In

line with strong rationalizability, player 2 must believe at h_2 that choosing b (that is, foregoing the outside option) is part of an optimal strategy for player 1. However, this is only possible if player 2 believes at h_2 that player 1 will choose f at h_3, as this is the only way for player 1 to obtain at least as much as by choosing a at the beginning.

As such, with the new information, the only strategy that player 2 can rationally choose in Γ is the strategy (d, g). However, this strategy cannot be rationally chosen in Γ with the original reasoning, where player 2 believes at h_2 that player 1 will choose e at h_3. As such, strong rationalizability *violates* supergame monotonicity.

Summarizing, we see that in the smaller game Γ, player 2 must believe at h_2 that player 1 will choose e, whereas in the supergame $\hat{\Gamma}$ player 2 must believe at h_2 that player 1 will choose f. That is, player 2's unique belief at h_2 is *overruled* by a new unique belief if player 2 learns about the outside option that player 1 could have chosen before the start of the game Γ. This contradicts supergame monotonicity.

8.2 Subgames and Supergames

To formally define supergame monotonicity, we first define what we mean by a subgame. Consider two dynamic games Γ and $\hat{\Gamma}$. Then, Γ is a *subgame* of $\hat{\Gamma}$ if (a) every player in Γ is also a player in $\hat{\Gamma}$, (b) every history in Γ is also a history in $\hat{\Gamma}$, (c) every terminal history in Γ is also a terminal history in $\hat{\Gamma}$, (d) every information set in Γ is also an information set in $\hat{\Gamma}$, (e) every information set in $\hat{\Gamma}$ that contains a history in Γ must also be an information set in Γ, (f) every action in $\hat{\Gamma}$ at an information set in Γ must also be an action in Γ and *vice versa*, and (g) for every player i in Γ, the utility function $\hat{u}_i$ in $\hat{\Gamma}$, when restricted to the terminal histories in Γ, coincides with i's utility function u_i in Γ.

In particular, condition (e) makes sure that a player in $\hat{\Gamma}$, when moving at a history in Γ, will know that the history at which he is moving is in Γ. As such, in the larger game $\hat{\Gamma}$ it is common knowledge between all players that are active in Γ, at each of the histories in Γ, that the current history is in Γ. Whenever Γ is a subgame of $\hat{\Gamma}$, we say that $\hat{\Gamma}$ is a *supergame* of Γ.

8.3 Supergame Monotonicity

Recall the intuition behind supergame monotonicity, which states that if a player learns that the game Γ was actually preceded by some earlier moves, resulting in a supergame $\hat{\Gamma}$, then this new information should only refine, but never overthrow, his previous reasoning. But what do we mean by the reasoning of a player in the games Γ and $\hat{\Gamma}$?

Formally, we model the output of the players' reasoning processes by conditional belief vectors, which specify at every information set where this player is active the belief that he holds about the other players' strategies. In general, a *solution concept* φ specifies for every game Γ and every player

i in Γ a set $B_i^{\varphi}(\Gamma)$ of conditional belief vectors, representing the possible beliefs that player i can end up with if he reasons according to the standards of φ.

Now suppose that the players in Γ learn that the game was actually preceded by earlier moves, resulting in the supergame $\hat{\Gamma}$. Then, the reasoning of the players in the new, larger game $\hat{\Gamma}$ is represented by the new sets of conditional belief vectors $B_i^{\varphi}(\hat{\Gamma})$. But what consequences does this new reasoning have for the choices in the smaller game Γ?

Let us denote by $S_i^{\hat{\Gamma}}$ and S_i^{Γ} the sets of strategies for player i in the games $\hat{\Gamma}$ and Γ, respectively. Similarly, we denote by $H_i^{\hat{\Gamma}}$ and H_i^{Γ} the collections of information sets in $\hat{\Gamma}$ and Γ, respectively, where player i is active.

Consider a conditional belief vector $\hat{b}_i \in B_i^{\varphi}(\hat{\Gamma})$ in the new game $\hat{\Gamma}$ allowed by the new reasoning, an information set $h \in H_i^{\Gamma}$ in the subgame, and a strategy $s_i \in S_i^{\Gamma}(h)$ for player i in the subgame that allows for h to be reached. Then, as we will argue below, we can define the expected utility $u_i(s_i, \hat{b}_i(h))$ induced by the strategy s_i in the subgame Γ and the conditional belief $\hat{b}_i(h)$ at h in the supergame $\hat{\Gamma}$.

To see this, consider a strategy $s_i \in S_i^{\Gamma}(h)$ in the subgame and an opponents' strategy combination $\hat{s}_{-i} \in S_{-i}^{\hat{\Gamma}}(h)$ in the supergame that allows for h to be reached. Then, $\hat{s}_{-i}$ induces an opponents' strategy combination in Γ, by restricting $\hat{s}_{-i}$ to the information sets in Γ. Consequently, $(s_i, \hat{s}_{-i})$ induces a terminal history $z(s_i, \hat{s}_{-i})$ in Γ.

Next, consider a strategy $s_i \in S_i^{\Gamma}(h)$ in the subgame that allows for h to be reached, and a conditional belief $\hat{b}_i(h)$ at h in the supergame $\hat{\Gamma}$. The expected utility at h, given s_i and $\hat{b}_i(h)$, is then given by

$$u_i(s_i, \hat{b}_i(h)) := \sum_{\hat{s}_{-i} \in S_{-i}^{\hat{\Gamma}}(h)} \hat{b}_i(h)(\hat{s}_{-i}) \cdot u_i(z(s_i, \hat{s}_{-i})).$$

We say that the strategy s_i is *optimal* for $\hat{b}_i$ at h if

$$u_i(s_i, \hat{b}_i(h)) \geq u_i(s_i', \hat{b}_i(h)) \text{ for all } s_i' \in S_i^{\Gamma}(h).$$

Then, we denote by

$$S_i^{\varphi}(\Gamma \mid \hat{\Gamma}) := \{s_i \in S_i^{\Gamma} \mid s_i \text{ is optimal for some } \hat{b}_i \in B_i^{\varphi}(\hat{\Gamma}) \text{ at all information sets } h \in H_i^{\Gamma}(s_i)\}$$

the set of strategies for player i that are optimal in the subgame Γ if the players learn that the actual game is $\hat{\Gamma}$.

We call $S_i^{\varphi}(\Gamma \mid \hat{\Gamma})$ the set of strategies that is *predicted for the subgame* Γ if the solution concept φ is applied to the supergame $\hat{\Gamma}$. In particular, $S_i^{\varphi}(\Gamma \mid \Gamma)$ contains those strategies that the solution concept induces for player i in the game Γ if the players do not learn any new information there.

Supergame monotonicity then states that every strategy in $S_i^{\varphi}(\Gamma \mid \hat{\Gamma})$, which is allowed by the new reasoning after learning that the actual game is $\hat{\Gamma}$, must also be allowed by the original reasoning which took place before receiving this new information – that is, it must be in $S_i^{\varphi}(\Gamma \mid \Gamma)$.

Definition 8.1 (Supergame monotonicity) *A solution concept φ satisfies supergame monotonicity if for every dynamic game Γ, every supergame $\hat{\Gamma}$, and every player i, it holds that $S_i^{\varphi}(\Gamma \mid \hat{\Gamma}) \subseteq S_i^{\varphi}(\Gamma \mid \Gamma)$.*

In general terms, supergame monotonicity thus states that every "solution" of the supergame, when restricted to the small game, should also be a "solution" of the small game itself. When read in this fashion, this property corresponds precisely to requirement BI1 in Kohlberg and Mertens (1986), which states that a solution of a game should always induce a solution in each of its subgames.

Kohlberg and Mertens use condition BI1 as a key characteristic of any backward induction concept. As such, supergame monotonicity can be viewed as a necessary condition for backward induction reasoning. We have seen above that the concept of strong rationalizability violates supergame monotonicity, which confirms that this concept is *not* based on backward induction reasoning.

8.4 Forward and Backward Rationalizability

It can be shown that the new concept proposed in this paper satisfies supergame monotonicity. In light of the discussion above, this indicates that the concept of forward and backward rationalizability is compatible with Kohlberg and Mertens' backward induction property BI1.

Theorem 8.1 (Supergame monotonicity) *The concept of forward and backward rationalizability satisfies supergame monotonicity.*

In Figure 1 let, as before, Γ be the game that starts at h_2 and $\hat{\Gamma}$ the supergame that starts at h_1. Then, by construction, the strategies $S_i^{\varphi}(\Gamma \mid \hat{\Gamma})$ that the forward and backward rationalizability procedure selects for the subgame Γ if the players learn that the actual game is $\hat{\Gamma}$ are the unique backward induction strategies in Γ. These, in turn, correspond to the strategies that are selected when the procedure is applied to Γ alone. As such, the reasoning of both players within Γ is not altered if we move from Γ to the supergame $\hat{\Gamma}$. In particular, supergame monotonicity holds for the concept of forward and backward rationalizability when moving from Γ to $\hat{\Gamma}$.

There are also examples where the additional information of past play, provided by the supergame $\hat{\Gamma}$, may *strictly refine* the reasoning of the players when using forward and backward rationalizability. Consider, for instance, Figure 2, where we focus on the game Γ that starts at h_3 and the supergame $\hat{\Gamma}$ starting at h_2. It is easily seen that forward and backward rationalizability allows for all possible choices in Γ if the concept is applied to Γ alone. If the players learn that the actual game is $\hat{\Gamma}$, then player 1 must believe at h_3 that player 2 chooses the strategy (c, h). Therefore, the concept selects for player 1 only the strategy f and for player 2 only the strategy h in Γ. Thus, the additional information about player 2's past play provided by the supergame $\hat{\Gamma}$ strictly refines the reasoning of player 1 within the original game Γ.

8.5 Relation with Strong Rationalizability Conditional on Subgames

When taken together, Theorems 6.3 and 8.1 imply the following property: Consider a game Γ embedded in a supergame $\hat{\Gamma}$. Suppose the players learn that the actual game being played is $\hat{\Gamma}$, and concentrate on the outcomes that the forward and backward rationalizability procedure selects for the subgame Γ. Then, every such outcome will also be possible if strong rationalizability is applied to the subgame Γ only. [13]

To see why this holds, let us denote by fbr the solution concept associated with forward and backward rationalizability. Take a strategy profile $(s_i)_{i \in I}$ in the subgame Γ, where $s_i \in S_i^{fbr}(\Gamma \mid \hat{\Gamma})$ for every player i, resulting in an outcome z in Γ. Since, by Theorem 8.1, our procedure satisfies supergame monotonicity, we must have that $s_i \in S_i^{fbr}(\Gamma \mid \Gamma)$ for every player i. Hence, the outcome z is also possible if fbr is applied to Γ only. But then, by Theorem 6.3 applied to the smaller game Γ, the outcome z must also be possible if strong rationalizability is applied to Γ only.

In the statement below, let fbr and sr refer to the forward and backward rationalizability concept, and the strong rationalizability concept, respectively.

Corollary 8.1 (Forward induction conditional on subgames) *Let the game Γ be embedded in a supergame $\hat{\Gamma}$. Then, $Z(S^{fbr}(\Gamma \mid \hat{\Gamma})) \subseteq Z(S^{sr}(\Gamma \mid \Gamma))$.*

That is, conditional on reaching a subgame our procedure refines, in terms of outcomes, the strong rationalizability procedure applied to this subgame alone. As an illustration, consider the double outside option game in Figure 2. Let $\hat{\Gamma}$ be the whole game, and Γ the subgame that starts at h_3. If we apply the forward and backward rationalizability procedure to $\hat{\Gamma}$, then the predicted strategies conditional on Γ would be f for player 1 and h for player 2. Hence, the unique predicted outcome conditional on Γ would be (f, h). On the other hand, if strong rationalizability would be applied to Γ alone, then every outcome in Γ would be possible.

8.6 Backwards Rationalizability

We now turn to the concept of backwards rationalizability, which provides an instance of pure backward induction reasoning. It turns out that this concept also satisfies supergame monotonicity.

Theorem 8.2 (Supergame monotonicity of backwards rationalizability) *The concept of backwards rationalizability satisfies supergame monotonicity.*

Since backwards rationalizability is a forward looking concept, one would be tempted to believe that the solution of the subgame Γ would not change if the players learn that the play has started at a supergame $\hat{\Gamma}$. That is, one may be led to think that $S^{br}(\Gamma \mid \hat{\Gamma}) = S^{br}(\Gamma \mid \Gamma)$ in this case, where br

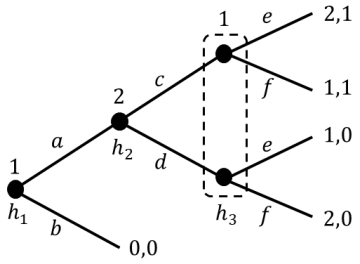

Figure 6: Supergame monotonicity does not hold with equality for backwards rationalizability

denotes the backwards rationalizability concept, such that supergame monotonicity would hold with equality. This, however, is not true, as the example in Figure 6 shows.

Let $\hat{\Gamma}$ be the supergame that starts at h_1, and Γ the subgame that starts at h_2. In Γ, without having the information that the game started at h_1, player 1 can rationally choose the strategies e and f under backwards rationalizability. Indeed, since player 2 moves before player 1, it follows that player 1's belief about player 2's strategy is unrestricted, and therefore player 1 can rationally choose e and f under backwards rationalizability. Hence, $S_1^{br}(\Gamma \mid \Gamma) = \{e, f\}$.

Suppose now that player 1 learns that the game started at $\hat{\Gamma}$. If we apply backwards rationalizability to $\hat{\Gamma}$, then player 1 must believe at h_1 that player 2 chooses the unique optimal strategy c at the future information set h_2. By Bayesian updating, player 1 must then assign probability 1 to player 2's strategy c at h_3. As a consequence, player 1's only optimal strategy under backwards rationalizability is (a, e). Therefore, $S_1^{br}(\Gamma \mid \hat{\Gamma}) = \{e\}$. We thus see that $S_1^{br}(\Gamma \mid \hat{\Gamma})$ is included in, but not equal to, $S_1^{br}(\Gamma \mid \Gamma)$. That is, supergame monotonicity does not hold with equality.

In other words, player 1's reasoning under backwards rationalizability is affected by the new information that the game has started at $\hat{\Gamma}$. However, it is only affected because of Bayesian updating, not because of strategic considerations. If we would drop the Bayesian updating condition from backwards rationalizability – resulting in the *backward dominance* procedure in Perea (2014) – then the weaker concept would satisfy supergame monotonicity with equality.

The stronger property, that supergame monotonicity is satisfied "with equality", corresponds to the combination of the backward induction properties BI1 and BI2 in Kohlberg and Mertens (1986). Their condition BI2 states that a solution of a subgame should also be part of a solution of the larger game it belongs to.

[13] Though strong rationalizability applied to the smaller game might also allow for more outcomes of the smaller game.

Kohlberg and Mertens insist on condition BI1, but reject condition BI2, as it precludes forward induction reasoning. This is precisely the reason why our concept of forward and backward rationalizability, which combines forward and backward induction reasoning, satisfies condition BI1 but violates condition BI2.

To see the latter, consider Figure 2, and let Γ be the game that starts at h_3. Then, forward and backward rationalizability allows for all possible choices there. Now, move to the supergame $\hat{\Gamma}$ that starts at h_1. Then, we have seen that the predicted strategies conditional on Γ only involve the choices f and h. As such, forward and backward rationalizability violates condition BI2, and therefore does not satisfy supergame monotonicity with equality.

The reason is clear: Condition BI2 states that when the player discovers that the actual game being played is the supergame $\hat{\Gamma}$, and not Γ, then the player should ignore the additional information provided by $\hat{\Gamma}$ in his reasoning process. This goes against the whole idea of forward induction which is present, for instance, in forward and backward rationalizability.

In Theorem 5.1 we have seen that forward and backward rationalizability provides a refinement of backwards rationalizability in terms of strategies. It is not hard to show that this result also holds conditional on every subgame. To see this, consider a subgame Γ within a supergame $\hat{\Gamma}$. Then, we know by Theorem 8.1 that every strategy $s_i \in S_i^{fbr}(\Gamma \mid \hat{\Gamma})$ is also in $s_i \in S_i^{fbr}(\Gamma \mid \Gamma)$. Moreover, by applying Theorem 5.1 to the smaller game Γ we know that $s_i \in S_i^{br}(\Gamma \mid \Gamma)$. That is, the strategy s_i is backwards rationalizable in the subgame Γ. We thus arrive at the conclusion below, where fbr and br refer to the forward and backward rationalizability concept, and the backwards rationalizability concept, respectively.

Corollary 8.2 (Backward induction conditional on subgames) *Let the game Γ be embedded in a supergame $\hat{\Gamma}$. Then, for every player i we have that $S_i^{fbr}(\Gamma \mid \hat{\Gamma}) \subseteq S_i^{br}(\Gamma \mid \Gamma)$.*

In view of Corollaries 8.1 and 8.2 we thus conclude that the forward and backward rationalizability procedure, conditional on every subgame, (a) provides a refinement, in terms of outcomes, of strong rationalizability applied to this subgame, and (b) provides a refinement, in terms of strategies, of backwards rationalizability applied to this subgame. In particular, forward and backward rationalizability, conditional on every subgame, will always weakly refine the outcomes that are induced by both strong rationalizability and backwards rationalizability. That is, we obtain the following result.

Corollary 8.3 *Let the game Γ be embedded in a supergame $\hat{\Gamma}$. Then,*

$$Z(S^{fbr}(\Gamma \mid \hat{\Gamma})) \subseteq Z(S^{br}(\Gamma \mid \Gamma)) \cap Z(S^{sr}(\Gamma \mid \Gamma)).$$

As a consequence, our concept will always be in line with forward and backward induction reasoning, even upon reaching unexpected subgames.

9 Discussion

Role of mistakes. On a conceptual level, an important difference between *forward and backward rationalizability* and *strong rationalizability* lies in the role of mistakes. In strong rationalizability, a player never believes his opponents to make mistakes in the execution of their strategies, only in the planning of their strategies. Indeed, when a player is surprised by another player's move, he believes the other player consciously made this move because he attributed a lower level of iterated strong belief in rationality to his opponents. In forward and backward rationalizability, a player typically explains surprising opponents' moves by believing that these are due to past mistakes in the execution of strategies, while believing that these opponents will reason and behave in accordance with *common* backward strong belief in rationality directly after the occurence of the mistake, and from then onwards. In this light, it would be interesting to embed both concepts within the framework of Battigalli and de Vito (2021), which explicitly distinguishes between plans and actual behavior.

Bounded rationality. The concept of forward and backward rationalizability can naturally be restricted to the part of the game that starts at a certain period, say period m. We would thus obtain a *bounded rationality* variant of the concept, where players do not actively reason about choices that were taken before period m. This may be important if players face cognitive limitations, which prevents them from reasoning about the entire game, but allows them to reason a limited number of periods ahead. Think, for instance, about end games in chess. Our epistemic characterization reveals that such a bounded rationality version of the concept is obtained if players reason in accordance with common backward strong belief in rationality from period m onwards. For strong rationalizability, on the other hand, such bounded rationality variants are absent as this concept requires the players, at all levels, to always reason about the entire game.

Computational tractability. In comparison with strong rationalizability, the procedure of forward and backward rationalizability is often easier to implement, especially when the dynamic game is large. The reason is that the latter procedure starts by applying the strong rationalizability at the last information sets, and then works its way backwards towards the beginning of the game. By doing so, it typically keeps the decision problems at these information sets rather small, which makes the computations lighter.

Invariance. Our concept is clearly not invariant, as it may prescribe different sets of strategies for dynamic games that share the same reduced normal form. A natural question that arises is whether we can find an invariant concept that shares the same philosophy as forward and backward rationalizability.

Supergame monotonicity. We have shown that forward and backward rationalizability satisfies supergame monotonicity, unlike strong rationalizability. The intuitive reason is that the former concept applies strong rationalizability to *every subgame*, in a backward inductive fashion. In that sense, the relation between these two concepts is analogous to the relation between Nash equilibrium and subgame perfect equilibrium.

10 Appendix

10.1 Proof of Section 3

To prove Theorem 3.1 we need the following result.

Lemma 10.1 (Strong belief of nested sets) *For a given player i let*

$$\emptyset \neq D_k \subseteq D_{k-1} \subseteq ... \subseteq D_0 = S_{-i}$$

be a sequence of nested subsets of opponents' strategy combinations. Then, there is a conditional belief vector $b_i \in B_i$ that strongly believes each of the sets $D_k, D_{k-1}, \ldots, D_0$.

Proof. Take some arbitrary probability distribution $p_i \in \Delta(S_{-i})$ such that $p_i(s_{-i}) > 0$ for all $s_{-i} \in S_{-i}$. For a given information set $h \in H_i$ let $m(h)$ be the highest number in $\{0, ..., k\}$ such that $S_{-i}(h) \cap D_{m(h)} \neq \emptyset$. Define the conditional belief $b_i(h)$ by

$$b_i(h)(s_{-i}) := \begin{cases} \dfrac{p_i(s_{-i})}{p_i(S_{-i}(h) \cap D_{m(h)})}, & \text{if } s_{-i} \in S_{-i}(h) \cap D_{m(h)} \\ 0, & \text{otherwise.} \end{cases} .$$

Then, it may be verified that the conditional belief vector $b_i = (b_i(h))_{h \in H_i}$ so constructed satisfies Bayesian updating, and strongly believes each of the sets $D_k, D_{k-1}, \ldots, D_0$. ■

Proof of Theorem 3.1. Recall that $B_i^{m.k}$ and $S_i^{m.k}$ are the sets of conditional belief vectors and strategies, respectively, that survive round k at period m of the forward and backward rationalizability procedure. We show, by induction on $m.k$, that $B_i^{m.k}$ and $S_i^{m.k}$ are always non-empty, starting with $M.0$.

By definition, $B_i^{M.0} = B_i$ and $S_i^{M.0} = S_i$. Applying Lemma 10.1 to the case $D_k = D_0 = S_{-i}$ shows that B_i is non-empty.

Now, suppose that m and k are such that $(m.k) \neq (M.0)$. We distinguish two cases: (1) $k \geq 1$, and (2) $k = 0$.

Case 1. Suppose that $k \geq 1$. Then, by definition, $B_i^{m.k} = \{b_i \in B_i^{m.k-1} \mid b_i \text{ strongly believes } S_{-i}^{m.k-1}\}$. By construction,

$$S_{-i}^{m.k-1} \subseteq S_{-i}^{m.k-2} \subseteq ... \subseteq S_{-i}^{M.0},$$

and $B_i^{m.k}$ consists of all those conditional belief vectors that strongly believe each of these nested sets. As, by the induction assumption, each of these nested sets is non-empty, it follows by Lemma 10.1 that $B_i^{m.k}$ is non-empty as well.

Now, take some $b_i \in B_i^{m.k}$. Then, b_i satisfies Bayesian updating. It follows from Lemma 8.13.2 in Perea (2012) that there is a strategy s_i that is optimal for b_i at all $h \in H_i(s_i)$. In particular, s_i is optimal for b_i from period m onwards, and hence s_i is in $S_i^{m,k}$. Thus, $S_i^{m.k}$ is non-empty.

Case 2. Suppose that $k = 0$. Then, by construction, $B_i^{m.0} = B_i^{m+1.K_{m+1}}$ and $S_i^{m.0} = S_i^{m+1.K_{m+1}}$, where K_{m+1} is the round at which the procedure at period $m+1$ terminates. By the induction assumption, $B_i^{m+1.K_{m+1}}$ and $S_i^{m+1.K_{m+1}}$ are non-empty, and thus $B_i^{m.0}$ and $S_i^{m.0}$ are non-empty as well.

It thus follows, by induction on m and k, that $S_i^{m.k}$ and $B_i^{m.k}$ are non-empty for every m and k, and for every player i. As the procedure terminates after finitely many rounds, we conclude that every player i has at least one strategy and one conditional belief vector that are forward and backward rationalizable. ■

10.2 Proofs of Section 4

Proof of Lemma 4.1. Consider some information set $h \in H_i$ of player i and let $s_i \in S_i(h)$ be a strategy that allows h to be reached. We first show that the set of types $T_i(s_i, h)$ of player i for whom playing s_i is optimal at h is a closed set of types. To this purpose, we consider for any alternative strategy $r_i \in S_i(h)$, any opponents' strategy combination s_{-i} and any opponents' type combination t_{-i} the utility difference

$$u_i(s_i, s_{-i}, t_{-i}) - u_i(r_i, s_{-i}, t_{-i}) := u_i(z(s_i, s_{-i})) - u_i(z(r_i, s_{-i})).$$

By fixing s_i, r_i, and h, and varying s_{-i} and t_{-i}, we obtain a bounded continuous function

$$u_i(s_i, \cdot, \cdot) - u_i(r_i, \cdot, \cdot) : S_{-i}(h) \times T_{-i} \to \mathbf{R}.$$

This is indeed the case, since we endow $S_{-i}(h)$ with the discrete topology and $u_i(s_i, s_{-i}, t_{-i}) - u_i(r_i, s_{-i}, t_{-i})$ only depends on the S_{-i}-dimension of $S_{-i} \times T_{-i}$. Hence, the set of measures $\mu_i \in \Delta(S_{-i}(h) \times T_{-i})$ such that integrating over this function with respect to μ_i is non-negative is a closed set of measures, which we call $\Delta(S_{-i}(h) \times T_{-i})^{s_i \geq r_i}$. The set of measures $\mu_i \in \Delta(S_{-i}(h) \times T_{-i})$ such that s_i is optimal at h is the intersection

$$\Delta(S_{-i}(h) \times T_{-i})^{s_i, h} := \bigcap_{r_i \in S_i(h)} \Delta(S_{-i}(h) \times T_{-i})^{s_i \geq r_i},$$

which is closed as well. Note that, by construction,

$$T_i(s_i, h) = \{t_i \in T_i \mid \beta_i(t_i, h) \in \Delta(S_{-i}(h) \times T_{-i})^{s_i, h}\}.$$

Since the mapping $\beta_i(\cdot, h) : T_i \to \Delta(S_{-i}(h) \times T_{-i})$ is continuous, we conclude that the set $T_i(s_i, h)$ is closed.

Recall that the set of types t_i such that s_i is optimal at h is precisely $T_i(s_i, h)$. For a given period m, let $H^{\geq m} := H^m \cup H^{m+1} \cup \ldots \cup H^M$ be the collection of information sets from period m onwards. Then, the set of types for which s_i is optimal from period m onwards is

$$T_i^{\geq m}(s_i) := \bigcap_{h \in H_i(s_i) \cap H^{\geq m}} T_i(s_i, h),$$

which is closed in T_i. Note that if s_i does not reach any information set in $H^{\geq m}$, then s_i is automatically optimal from period m onwards for all types in T_i. For each of the finitely many strategies $s_i \in S_i$ of player i, the set $\{s_i\} \times T_i^{\geq m}(s_i)$ is closed in the product topology of $S_i \times T_i$, since it is the product of two closed sets. The set

$$(S_i \times T_i)^{rat,m} = \bigcup_{s_i \in S_i} (\{s_i\} \times T_i^{\geq m}(s_i))$$

is closed in $S_i \times T_i$ since it is the union of finitely many closed sets. If $\hat{T}_i$ is a measurable subset of T_i then $(S_i \times \hat{T}_i)^{rat,m} = (S_i \times T_i)^{rat,m} \cap (S_i \times \hat{T}_i)$ is measurable since it is an intersection of a closed and a measurable set. If $\hat{T}_i$ is closed, then $S_i \times \hat{T}_i$ is closed and hence $(S_i \times \hat{T}_i)^{rat,m}$ is closed, being the intersection of two closed sets. ■

Proof of Lemma 4.2. We start by proving the following result.

Claim. Let E be a closed subset of $S_{-i} \times T_{-i}$. Then, the set $\{ t_i \mid t_i$ strongly believes $E \}$ is a closed subset of T_i.

Proof of claim. Let $h \in H_i$ be such that $(S_{-i}(h) \times T_{-i}) \cap E \neq \emptyset$. We show that the set of measures in $\Delta(S_{-i}(h) \times T_{-i})$ that assign probability 1 to E is closed set. To this end, let $(\mu_n)_{n \in \mathbb{N}} \to \mu$ be a sequence of probability measures in $\Delta(S_{-i}(h) \times T_{-i})$ converging to $\mu \in \Delta(S_{-i}(h) \times T_{-i})$ such that $\mu_n(E) = 1$ for all $n \in \mathbb{N}$. We have to show that $\mu(E) = 1$. But this follows immediately from the Portemanteau Theorem (Kechris (1995), Theorem 17.20). By continuity, the set $\{t_i \in T_i \mid \beta_i(t_i, h)(E) = 1\}$ is a closed set of types. The set of types t_i that strongly believe E is the finite intersection of such sets of types over all $h \in H_i$ such that $(S_{-i}(h) \times T_{-i}) \cap E \neq \emptyset$. Hence, this is a closed set of types. ◊

The lemma now follows immediately by iteratively applying the claim and Lemma 4.1. ■

Proof of Theorem 4.1. As a first step we will use the forward and backward rationalizability procedure to build a finite type space. Later we will use this model to prove the theorem. Moreover, we will make sure that the type space is non-reduntant, that is, no two different types of a player induce the same conditional belief hierarchy.

Recall that, for every player i, period m, and round k, the sets $B_i^{m.k}$ and $S_i^{m.k}$ are the collections of conditional belief vectors and strategies, respectively, selected by the forward and backward rationalizability procedure at round k of period m. In particular, $B_i^{L.K_L}$ and $S_i^{L.K_L}$ are the sets of conditional belief hierarchies and strategies, respectively, that survive all rounds at all periods.

For every player i and strategy $s_i \in S_i^{L.K_L}$ choose a conditional belief vector $b_i[s_i] \in B_i^{L.K_L}$ such that s_i is optimal for $b_i[s_i]$ from the first period onwards.

For all other strategies s_i there is a period $m \in \{L, ..., M\}$ and a round k such that $s_i \in S_i^{m.k} \backslash S_i^{m.k+1}$. For such a strategy $s_i \in S_i^{m.k} \backslash S_i^{m.k+1}$ we can then choose a conditional belief vector $b_i[s_i] \in B_i^{m.k}$ such that s_i is optimal for $b_i[s_i]$ from period m onwards if $k \geq 1$, and s_i is optimal

from period $m+1$ onwards if $k=0$. If $m.k = M.0$, then optimality from period $M+1$ onwards means that s_i need not be optimal for $b_i[s_i]$ at all.

Based on these conditional belief vectors $b_i[s_i]$ we will now construct a finite type space $\hat{\mathcal{T}} = ((T_i, \mathcal{O}_i), \beta_i)$ where the sets of types are given by $T_i = \{t_i^{b_i[s_i]} \mid s_i \in S_i\}$, and the belief mappings β_i are such that

$$\beta_i(t_i^{b_i[s_i]}, h)((s_j, t_j)_{j \neq i}) = \begin{cases} b_i[s_i](h)((s_j)_{j \neq i}), & \text{if } t_j = t_j^{b_j[s_j]} \text{ for all } j \neq i \\ 0, & \text{otherwise} \end{cases} \tag{10.1}$$

for all players i, all strategies s_i, all information sets $h \in H_i$, and all opponents' strategy-type combinations $(s_j, t_j)_{j \neq i} \in S_{-i} \times T_{-i}$. Hence, every type $t_i^{b_i[s_i]}$ has the belief $b_i[s_i](h)$ about the opponents' strategy combinations at every information set $h \in H_i$, and matches, in its belief, every opponent's strategy s_j with the associated type $t_j^{b_j[s_j]}$. It is easy to see that every type in this model satisfies Bayesian updating. Note that $b_i[s_i] = b_i[\hat{s}_i]$ implies that $t_i^{b_i[s_i]} = t_i^{b_i[\hat{s}_i]}$, and hence the type space $\hat{\mathcal{T}}$ is non-redundant by construction.

For every player i and conditional belief vector $b_i \in B_i \backslash \{b_i[s_i] \mid s_i \in S_i\}$ not present in $\hat{\mathcal{T}}$, we add a new type $t_i^{b_i}$ to $\hat{\mathcal{T}}$ whose conditional beliefs are given by

$$\beta_i(t_i^{b_i}, h)((s_j, t_j)_{j \neq i}) = \begin{cases} b_i(h)((s_j)_{j \neq i}), & \text{if } t_j = t_j^{b_j[s_j]} \text{ for all } j \neq i \\ 0, & \text{otherwise} \end{cases} \tag{10.2}$$

The new type space obtained after adding the type $t_i^{b_i}$ to $\hat{\mathcal{T}}$ is denoted by $\hat{\mathcal{T}} \cup \{t_i^{b_i}\}$.

Let $\mathcal{T}$ be a universal type space. Then, by definition, each of the finite type spaces $\hat{\mathcal{T}}$ and $\hat{\mathcal{T}} \cup \{t_i^{b_i}\}$ maps in a unique way to the universal type space $\mathcal{T}$ by a type morphism. Note that for every type $t_i^{b_i[s_i]}$ in $\hat{\mathcal{T}}$ the induced conditional belief hierarchy is the same, no matter whether it is regarded as a type in $\hat{\mathcal{T}}$ or a type in $\hat{\mathcal{T}} \cup \{t_j^{bj}\}$. Since a type morphism always preserves the induced conditional belief hierarchy, the type $t_i^{b_i[s_i]}$ will be mapped to the same type in the universal type space $\mathcal{T}$, no matter whether it is regarded as a type in $\hat{\mathcal{T}}$ or a type in $\hat{\mathcal{T}} \cup \{t_j^{bj}\}$. As each of the type spaces $\hat{\mathcal{T}}$ and $\hat{\mathcal{T}} \cup \{t_i^{b_i}\}$ is non-redundant, every type in these type spaces may be uniquely identified with a type in the universal type space $\mathcal{T}$.

For every player i, period m and number $k \in \{0, 1, ...\}$, we denote by $T_i^{m.k}$ the set of types for player i in the universal type space $\mathcal{T}$ that express k-fold backward strong belief in rationality from period m onwards. Define

$$\hat{B}_i^{m.k} := \{b_i \in B_i \mid \text{there is some } t_i \in T_i^{m.k} \text{ that induces the conditional belief vector } b_i\}$$

and

$$\hat{S}_i^{m.k} := \{s_i \in S_i \mid \text{there is some } t_i \in T_i^{m.k} \text{ with } (s_i, t_i) \in (S_i \times T_i^{m.k})^{rat,m}\}.$$

Here, when we say that "t_i induces the conditional belief vector b_i", we mean that $\text{marg}_{S_{-i}(h)}\beta_i(t_i, h) = b_i(h)$ for every $h \in H_i$. We prove the following result.

Claim. For every period m and number $k \in \{0, 1, ...\}$ it holds that (i) $\hat{B}_i^{m.k} \subseteq B_i^{m.k+1}$, (ii) $B_i^{m.k+1} \subseteq \hat{B}_i^{m.k}$ and for every $b_i \in B_i^{m.k+1}$ we have that $t_i^{b_i} \in T_i^{m.k}$, (iii) $\hat{S}_i^{m.k} \subseteq S_i^{m.k+1}$ and (iv) $S_i^{m.k+1} \subseteq \hat{S}_i^{m.k}$.

Proof of claim. We show the four statements by induction on $m.k$.

We start with $M.0$. Then, $\hat{B}_i^{M.0}$ is, by definition, the set of conditional belief vectors induced by the types in $T_i^{M.0}$. As $T_i^{M.0} = T_i$, this is the set of all conditional belief vectors, and hence $\hat{B}_i^{M.0} = B_i$. As, by construction, $B_i^{M.1} = B_i$ as well, it follows that $\hat{B}_i^{M.0} = B_i^{M.1}$. Moreover, as $T_i^{M.0} = T_i$, for every $b_i \in B_i^{M.1}$ we have that $t_i^{b_i} \in T_i^{M.0}$. This establishes (i) and (ii).

Moreover, $\hat{S}_i^{M.0}$ contains precisely those strategies that are optimal from period M onwards for some $t_i \in T_i^{M.0}$. As $T_i^{M.0} = T_i$, these are precisely the strategies that are optimal from period M onwards for some conditional belief vector $b_i \in B_i$. By definition, these are precisely the strategies in $S_i^{M.1}$. Hence, we conclude that $\hat{S}_i^{M.0} = S_i^{M.1}$. This establishes (iii) and (iv).

Next, take some $m.k \neq M.0$, and assume that the claim holds for $m.k-1$ if $k \geq 1$, and that the claim holds for any $m+1.k'$ if $k = 0$. We distinguish two cases: (1) $k = 0$, and (2) $k \geq 1$.

Case 1. Suppose that $k = 0$. Then, by definition, there is some round K such that $\hat{B}_i^{m.0} = \hat{B}_i^{m+1.K}$ and $B_i^{m.1} = B_i^{m+1.K+1}$. As, by the induction assumption, $\hat{B}_i^{m+1.K} = B_i^{m+1.K+1}$, we conclude that $\hat{B}_i^{m.0} = B_i^{m.1}$. Moreover, by construction, $\hat{S}_i^{m.0}$ is the set of strategies that are optimal, from period m onwards, for some $b_i \in \hat{B}_i^{m.0}$, whereas $S_i^{m.1}$ is the set of strategies that are optimal, from period m onwards, for some $b_i \in B_i^{m.1}$. Since $\hat{B}_i^{m.0} = B_i^{m.1}$, it follows that $\hat{S}_i^{m.0} = S_i^{m.1}$.

Case 2. Suppose that $k \geq 1$.

(i) We show that $\hat{B}_i^{m.k} \subseteq B_i^{m.k+1}$. Take some $b_i \in \hat{B}_i^{m.k}$. Then, there is some $t_i \in T_i^{m.k}$ that induces b_i. By definition, $T_i^{m.k} \subseteq T_i^{m.k-1}$, and hence $b_i \in \hat{B}_i^{m.k-1}$. By the induction assumption on (i) it follows that $b_i \in B_i^{m.k}$. Hence, we only need to show that b_i strongly believes $S_{-i}^{m.k}$. Let $h \in H_i$ be such that $S_{-i}^{m.k} \cap S_{-i}(h) \neq \emptyset$. We must show that $b_i(h)(S_{-i}^{m.k}) = 1$. By the induction assumption applied to (iii) and (iv) we know that $S_{-i}^{m.k} = \hat{S}_{-i}^{m.k-1}$. Hence, by the definition of $\hat{S}_j^{m.k-1}$ for every $j \neq i$, we know that $\times_{j \neq i}(S_j \times T_j^{m.k-1})^{rat,m} \cap (S_{-i}(h) \times T_{-i}) \neq \emptyset$. Therefore, since $t_i \in T_i^{m.k}$, we conclude that $\beta_i(t_i, h)(\times_{j \neq i}(S_j \times T_j^{m.k-1})^{rat,m}) = 1$. This, in turn, implies that $b_i(h)(\hat{S}_{-i}^{m.k-1}) = 1$. As, by the induction assumption on (iii) and (iv), $S_{-i}^{m.k} = \hat{S}_{-i}^{m.k-1}$, we conclude that $b_i(h)(S_{-i}^{m.k}) = 1$. Hence, b_i strongly believes $S_{-i}^{m.k}$. Since $b_i \in B_i^{m.k}$, it follows that $b_i \in B_i^{m.k+1}$. As such, $\hat{B}_i^{m.k} \subseteq B_i^{m.k+1}$.

(ii) We show that $B_i^{m.k+1} \subseteq \hat{B}_i^{m.k}$ and for every $b_i \in B_i^{m.k+1}$ we have that $t_i^{b_i} \in T_i^{m.k}$. Take some $b_i \in B_i^{m.k+1}$. Then, in particular, $b_i \in B_i^{m.k}$ and hence we know, by the induction assumption on (ii), that $t_i^{b_i} \in T_i^{m.k-1}$. Thus, to prove that $t_i^{b_i} \in T_i^{m.k}$ it only remains to show that $t_i^{b_i}$ strongly believes $\times_{j \neq i}(S_j \times T_j^{m.k-1})^{rat.m}$. To this end, let $h \in H_i$ be such that $(S_{-i}(h) \times T_{-i}) \cap (\times_{j \neq i}(S_j \times T_j^{m.k-1})^{rat,m}) \neq \emptyset$. We must show that $\beta_i(t_i^{b_i}, h)(\times_{j \neq i}(S_j \times T_j^{m.k-1})^{rat,m}) = 1$.

By definition, $\hat{S}_{-i}^{m.k-1} = \text{proj}_{S_{-i}}(\times_{j \neq i}(S_j \times T_j^{m.k-1})^{rat,m})$. Note that by the induction assumption of (iii) and (iv) we have that $\hat{S}_{-i}^{m.k-1} = S_{-i}^{m.k}$. Therefore, we have $S_{-i}(h) \cap S_{-i}^{m.k} \neq \emptyset$. Since $b_i \in B_i^{m.k+1}$ it follows that $b_i(h)(S_{-i}^{m.k}) = 1$, and hence $b_i(h)(\hat{S}_{-i}^{m.k-1}) = 1$. By the definition of $t_i^{b_i}$ in (10.2) we

have $\beta_i(t_i^{b_i}, h)((\hat{S}_{-i}^{m.k-1} \cap S_{-i}(h)) \times T_{-i}) = 1$, and that $\beta_i(t_i, h)$ assigns probability 1 to the set of opponents' strategy-type combinations $\{(s_j, t_j^{b_j[s_j]})_{j \neq i} \mid s_j \in \hat{S}_j^{m.k-1} \text{ for all } j \neq i\}$. As $\hat{S}_{-i}^{m.k-1} = S_{-i}^{m.k}$ we conclude that $\beta_i(t_i^{b_i}, h)$ assigns probability 1 to the set of opponents' strategy-type combinations $\{(s_j, t_j^{b_j[s_j]})_{j \neq i} \mid s_j \in S_j^{m.k} \text{ for all } j \neq i\}$.

Consider a type $t_j^{b_j[s_j]}$ where $s_j \in S_j^{m.k}$. Then, we know by the definition of type $t_j^{b_j[s_j]}$ in (10.1) that $t_j^{b_j[s_j]}$ induces the conditional belief vector $b_j[s_j] \in B_j^{m.k}$, and that s_j is optimal for $b_j[s_j]$ from period m onwards. Hence, s_j is optimal for $t_j^{b_j[s_j]}$ from period m onwards. As $b_j[s_j] \in B_j^{m.k}$ we conclude by the induction assumption of (ii) that $t_j^{b_j[s_j]} \in T_j^{m.k-1}$. As s_j is optimal for type $t_j^{b_j[s_j]}$ from period m onwards, it follows that $(s_j, t_j^{b_j[s_j]}) \in (S_j \times T_j^{m.k-1})^{rat,m}$. Recall that $\beta_i(t_i^{b_i}, h)$ assigns probability 1 to the set of opponents' strategy-type combinations $\{(s_j, t_j^{b_j[s_j]})_{j \neq i} \mid s_j \in S_j^{m.k} \text{ for all } j \neq i\}$. Hence, it follows that $\beta_i(t_i^{b_i}, h)(\times_{j \neq i}(S_j \times T_j^{m.k-1})^{rat,m}) = 1$. As such, we conclude that $t_i^{b_i}$ strongly believes $\times_{j \neq i}(S_j \times T_j^{m.k-1})^{rat.m}$.

Since $t_i^{b_i} \in T_i^{m.k-1}$ it follows that $t_i^{b_i} \in T_i^{m.k}$. We thus conclude that for every $b_i \in B_i^{m.k+1}$ we have that $t_i^{b_i} \in T_i^{m.k}$. Since, by (10.2), $t_i^{b_i}$ induces the conditional belief vector b_i, it follows that $b_i \in \hat{B}_i^{m.k}$. Hence, $B_i^{m.k+1} \subseteq \hat{B}_i^{m.k}$.

(iii) We show that $\hat{S}_i^{m.k} \subseteq S_i^{m.k+1}$. Let $s_i \in \hat{S}_i^{m.k}$. Then, in particular, $s_i \in \hat{S}_i^{m.k-1}$. By the induction assumption of (iii) it follows that $s_i \in S_i^{m.k}$. Since $s_i \in \hat{S}_i^{m.k}$, there is a $t_i \in T_i^{m.k}$ such that s_i is optimal for t_i from period m onwards. Let b_i be the conditional belief vector induced by t_i. As the expected utility depends only on first-order beliefs, s_i is optimal for b_i from period m onwards. Since $t_i \in T_i^{m.k}$ it follows, by definition, that $b_i \in \hat{B}_i^{m.k}$. By (i) it then follows that $b_i \in B_i^{m.k+1}$. Hence, s_i is optimal for some $b_i \in B_i^{m.k+1}$ from period m onwards. As we have seen above that $s_i \in S_i^{m.k}$, we conclude that $s_i \in S_i^{m.k+1}$. Hence, $\hat{S}_i^{m.k} \subseteq S_i^{m.k+1}$.

(iv) We finally show that $S_i^{m.k+1} \subseteq \hat{S}_i^{m.k}$. Let $s_i \in S_i^{m.k+1}$. Then, by construction, $b_i[s_i] \in B_i^{m.k+1}$ and s_i is optimal for $b_i[s_i]$ from period m onwards. By (ii) we know that $t_i^{b_i[s_i]} \in T_i^{m.k}$. Moreover, $t_i^{b_i[s_i]}$ induces the conditional belief vector $b_i[s_i]$. Since the expected utility depends only on first-order beliefs, we conclude that s_i is optimal for $t_i^{b_i[s_i]}$ from period m onwards. This implies that $(s_i, t_i^{b_i[s_i]}) \in (S_i \times T_i^{m.k})^{rat.m}$, and hence $s_i \in \hat{S}_i^{m.k}$. Thus, $S_i^{m.k+1} \subseteq \hat{S}_i^{m.k}$. This completes the proof of the claim. $\Diamond$

We are now able to prove the theorem.

(a) Take first a strategy s_i that is forward and backward rationalizable. Then, there is a conditional belief vector $b_i \in B_i^{L.K_L}$ such that s_i is optimal for b_i from the first period onwards. Note that $b_i \in B_i^{L.k+1}$ for all k and hence, by part (ii) of the claim, $t_i^{b_i} \in T_i^{L.k}$ for all k. Therefore, $t_i^{b_i} \in T_i^L$, and hence $t_i^{b_i}$ expresses common backward strong belief in rationality. As $t_i^{b_i}$ induces the conditional belief vector b_i, and s_i is optimal for b_i from the first period onwards, it follows that s_i is optimal for

$t_i^{b_i}$ from the first period onwards. As such, s_i is optimal, from the first period onwards, for a type that expresses common backward strong belief in rationality.

Conversely, suppose that s_i is optimal, from the the first period onwards, for a type t_i that expresses common backward strong belief in rationality. Hence, $t_i \in T_i^L$. Suppose that t_i induces the conditional belief vector b_i. Then, s_i is optimal, from the first period onwards, for b_i. Since $t_i \in T_i^{L.k}$ for all k, and t_i induces the conditional belief vector b_i, it follows that $b_i \in \hat{B}_i^{L.k}$ for all k. By part (i) of the claim it follows that $b_i \in B_i^{L.k+1}$ for all k, and hence b_i is forward and backward rationalizable. Since s_i is optimal for b_i from the first period onwards, we conclude that s_i is forward and backward rationalizable.

(b) Take first a strategy $s_i \in S_i^{m.0}$. Then, $s_i \in S_i^{m+1.K_{m+1}}$. Hence, there is a conditional belief vector $b_i \in B_i^{m+1.K_{m+1}}$ such that s_i is optimal for b_i from period $m+1$ onwards. Note that $b_i \in B_i^{m+1.k+1}$ for all k and hence, by part (ii) of the claim, $t_i^{b_i} \in T_i^{m+1.k}$ for all k. Therefore, $t_i^{b_i} \in T_i^{m+1}$, and hence $t_i^{b_i}$ expresses common backward strong belief in rationality from period $m+1$ onwards. As $t_i^{b_i}$ induces the conditional belief vector b_i, and s_i is optimal for b_i from period $m+1$ onwards, it follows that s_i is optimal for $t_i^{b_i}$ from period $m+1$ onwards. As such, s_i is optimal, from period $m+1$ onwards, for a type that expresses common backward strong belief in rationality from period $m+1$ onwards.

Conversely, suppose that s_i is optimal, from period $m+1$ onwards, for a type t_i that expresses common backward strong belief in rationality from period $m+1$ onwards. Hence, $t_i \in T_i^{m+1}$. Suppose that t_i induces the conditional belief vector b_i. Then, s_i is optimal, from period $m+1$ onwards, for b_i. Since $t_i \in T_i^{m+1.k}$ for all k, and t_i induces the conditional belief vector b_i, it follows that $b_i \in \hat{B}_i^{m+1.k}$ for all k. By part (i) of the claim it follows that $b_i \in B_i^{m+1.k+1}$ for all k, and hence $b_i \in B_i^{m.0}$. Since s_i is optimal for b_i from period $m+1$ onwards, we conclude that $s_i \in S_i^{m.0}$.

(c) Take first a strategy $s_i \in S_i^{m.k+1}$. Hence, there is a conditional belief vector $b_i \in B_i^{m.k+1}$ such that s_i is optimal for b_i from period m onwards. By part (ii) of the claim we conclude that $t_i^{b_i} \in T_i^{m.k}$. As $t_i^{b_i}$ induces the conditional belief vector b_i, and s_i is optimal for b_i from period m onwards, it follows that s_i is optimal for $t_i^{b_i}$ from period m onwards. As such, s_i is optimal, from period m onwards, for a type in $T_i^{m.k}$ that expresses k-fold backward strong belief in rationality from period m onwards.

Conversely, suppose that s_i is optimal, from period m onwards, for a type $t_i \in T_i^{m.k}$ that expresses k-fold backward strong belief in rationality from period m onwards. Suppose that t_i induces the conditional belief vector b_i. Then, s_i is optimal, from period m onwards, for b_i. Since $t_i \in T_i^{m.k}$ and t_i induces the conditional belief vector b_i, it follows that $b_i \in \hat{B}_i^{m.k}$. By part (i) of the claim it follows that $b_i \in B_i^{m.k+1}$. Since s_i is optimal for b_i from period m onwards, we conclude that $s_i \in S_i^{m.k+1}$.

This completes the proof. ■

10.3 Proof of Section 5

Proof of Theorem 5.1. Let $S_i^{fbr,m.k}$ and $B_i^{fbr,m.k}$ be the sets of strategies and conditional belief vectors for player i that result from period m, round k, of the forward and backward rationalizability

procedure. Similarly, we define the sets $S_i^{br,m.k}$ and $B_i^{br,m.k}$ for the backwards rationalizability procedure. For every period m, let K_m be the first round where both procedures terminate at period m. We show, by induction on $m.k$, that for all players i,

$$B_i^{fbr,m.k} \subseteq B_i^{br,m.k} \text{ and } S_i^{fbr,m.k} \subseteq S_i^{br,m.k}. \tag{10.3}$$

For $m.k = M.0$ this is true, since $B_i^{fbr,m.k} = B_i^{br,m.k} = B_i$ and $S_i^{fbr,m.k} = S_i^{br,m.k} = S_i$.

Now, take some $m.k \neq M.0$, and assume that (10.3) holds for $m+1.K_{m+1}$ if $k = 0$ or for $m.k-1$ if $k \geq 1$. If $k = 0$ then, by definition, $B_i^{fbr,m.0} = B_i^{fbr,m+1.K_{m+1}}$, $B_i^{br,m.0} = B_i^{br,m+1.K_{m+1}}$, $S_i^{fbr,m.0} = S_i^{fbr,m+1.K_{m+1}}$ and $S_i^{br,m.0} = S_i^{br,m+1.K_{m+1}}$. Thus, (10.3) would hold trivially, by the induction assumption.

Suppose now that $k \geq 1$ and that (10.3) holds for $m.k-1$. We first show that $B_i^{fbr,m.k} \subseteq B_i^{br,m.k}$. Take some $b_i \in B_i^{fbr,m.k}$. Then, in particular, $b_i \in B_i^{fbr,m.k-1}$. By the induction assumption it follows that $b_i \in B_i^{br,m.k-1}$. As $b_i \in B_i^{fbr,m.k}$, we know that b_i strongly believes $S_{-i}^{fbr,m.k-1}$. Take some $h \in H_i^m$. Since so far no restrictions have been imposed on the optimality of strategies at information sets preceding h, strongly believing $S_{-i}^{fbr,m.k-1}$ implies that $b_i(h)(S_{-i}^{fbr,m.k-1}) = 1$. As, by the induction assumption, $S_{-i}^{fbr,m.k-1} \subseteq S_{-i}^{br,m.k-1}$, it follows that $b_i(h)(S_{-i}^{br,m.k-1}) = 1$. As $b_i \in B_i^{br,m.k-1}$, we conclude that $b_i \in B_i^{br,m.k}$. Hence, $B_i^{fbr,m.k} \subseteq B_i^{br,m.k}$.

We next show that $S_i^{fbr,m.k} \subseteq S_i^{br,m.k}$. Take some $s_i \in S_i^{fbr,m.k}$. Then, s_i is optimal, from period m onwards, for some $b_i \in B_i^{fbr,m.k}$. As we have seen above that $B_i^{fbr,m.k} \subseteq B_i^{br,m.k}$, we conclude that s_i is optimal, from period m onwards, for some $b_i \in B_i^{br,m.k}$. Hence, $s_i \in S_i^{br,m.k}$. We have thus shown that $S_i^{fbr,m.k} \subseteq S_i^{br,m.k}$.

By induction, it follows that $B_i^{fbr,L.K_L} \subseteq B_i^{br,L.K_L}$ and $S_i^{fbr,L.K_L} \subseteq S_i^{br,L.K_L}$, which completes the proof. ■

10.4 Proof of Section 6

Proof of Lemma 6.1. Without loss of generality, suppose that the first time period is 1. For every period $m \in \{1, ..., M\}$ and $k \in \{0, 1, 2, ..., K_m\}$, let $m.k$ be the elimination step in period m, round k, of the forward and backward rationalizability procedure. Here, K_m denotes the last round of the procedure from period m onwards. This leads to the sequence

$$(D^{M.0}, D^{M.1}, ..., D^{M.K_M},\ D^{M-1.0}, D^{M-1.1}, ..., D^{M-1.K_{M-1}},\ ...\ , D^{1.0}, D^{1.1}, ..., D^{1,K_1})$$

of products of strategy sets.

We show that this sequence is an elimination order for sb. By definition, $D^{M.0} = \times_{i \in I} S_i$ and $sb(D^{1.K_1}) = D^{1.K_1}$. It remains to show condition (b) in the definition of an elimination order.

Consider first a step $m.k$ with $k \leq K_m - 1$. Then, by definition, $D^{m.k+1} = \times_{i \in I} D_i^{m.k+1}$, where

$$D_i^{m.k+1} = \{s_i \in D_i^{m.k} \mid \text{there is some } b_i \in B_i \text{ that strongly believes } D_{-i}^{M.0}, D_{-i}^{M.1}, ..., D_{-i}^{m.k} \text{ such that } s_i \text{ is optimal for } b_i \text{ from period } m \text{ onwards}\} \tag{10.4}$$

for every player i. Define, for every player i, the set

$$E_i^{m.k+1} = \{s_i \in D_i^{m.k} \mid \text{there is some } b_i \in B_i \text{ that strongly believes } D_{-i}^{m.k} \text{ such that } s_i \text{ is optimal for } b_i \text{ at every } h \in H(D^{m.k}) \cap H_i(s_i) \text{ that belongs to } H^\tau \text{ for some } \tau \geq m\}. \quad (10.5)$$

Claim. $E_i^{m.k+1} = D_i^{m.k+1}$.

Proof of claim. Clearly, $D_i^{m.k+1} \subseteq E_i^{m.k+1}$. To prove that $E_i^{m.k+1} \subseteq D_i^{m.k+1}$, take some $s_i \in E_i^{m.k+1}$. We will show that $s_i \in D_i^{m.k+1}$. As $s_i \in E_i^{m.k+1}$, there is some $b_i^{m.k} \in B_i$ that strongly believes $D_{-i}^{m.k}$ such that s_i is optimal for $b_i^{m.k}$ at every $h \in H(D^{m.k}) \cap H_i(s_i)$ that belongs to H^τ for some $\tau \geq m$. To show that $s_i \in D_i^{m.k+1}$ we distinguish two cases: (1) $k \geq 1$ and (2) $k = 0$.

Case 1. Suppose that $k \geq 1$. As $s_i \in E_i^{m.k+1}$ we know, by definition, that $s_i \in D_i^{m.k}$. Hence, there is some $b_i' \in B_i$ that strongly believes $D_{-i}^{M.0}, ..., D_{-i}^{m.k-1}$ such that s_i is optimal for b_i' from period m onwards. Define the conditional belief vector b_i where

$$b_i(h) := \begin{cases} b_i^{m.k}(h), & \text{if } D_{-i}^{m.k} \cap S_{-i}(h) \neq \emptyset \\ b_i'(h), & \text{otherwise} \end{cases}$$

for every $h \in H_i$.

Then, by construction, b_i strongly believes $D_{-i}^{M.0}, D_{-i}^{M.1}, ..., D_{-i}^{m.k}$. To show that s_i is optimal for b_i from period m onwards, take some $h \in H_i(s_i) \cap H^\tau$ for some $\tau \geq m$. We distinguish two cases.

If $D_{-i}^{m.k} \cap S_{-i}(h) \neq \emptyset$, then $h \in H(D^{m.k}) \cap H_i(s_i)$ since $s_i \in D_i^{m.k}$. Moreover, $b_i(h) = b_i^{m.k}(h)$. As, by construction, s_i is optimal for $b_i^{m.k}$ at h, it follows that s_i is optimal for b_i at h.

If $D_{-i}^{m.k} \cap S_{-i}(h) = \emptyset$ then, by construction, $b_i(h) = b_i'(h)$. As s_i is optimal for b_i' at h, it follows that s_i is optimal for b_i at h. Hence, s_i is optimal for b_i from period m onwards.

We finally show that b_i is in B_i, by proving that it satisfies Bayesian updating. Consider some information sets $h, h' \in H_i$, where h precedes h'. We distinguish three cases:

(i) If $D_{-i}^{m.k} \cap S_{-i}(h) \neq \emptyset$ and $D_{-i}^{m.k} \cap S_{-i}(h') \neq \emptyset$, then b_i coincides with $b_i^{m.k} \in B_i$ at h and h'. Since $b_i^{m.k}$ satisfies Bayesian updating, it follows that b_i satisfies Bayesian updating if the game moves from h to h'.

(ii) If $D_{-i}^{m.k} \cap S_{-i}(h) = \emptyset$ and $D_{-i}^{m.k} \cap S_{-i}(h') = \emptyset$, then b_i coincides with $b_i' \in B_i$ at h and h'. Since b_i' satisfies Bayesian updating, it follows that b_i satisfies Bayesian updating if the game moves from h to h'.

(iii) Suppose, finally, that $D_{-i}^{m.k} \cap S_{-i}(h) \neq \emptyset$ and $D_{-i}^{m.k} \cap S_{-i}(h') = \emptyset$. Then, $b_i(h) = b_i^{m.k}(h)$ and $b_i(h') = b_i'(h')$. As $D_{-i}^{m.k} \cap S_{-i}(h') = \emptyset$ and $b_i^{m.k}$ strongly believes $D_{-i}^{m.k}$, we have that $b_i^{m.k}(h)(D_{-i}^{m.k}) = 1$. Since $D_{-i}^{m.k} \cap S_{-i}(h') = \emptyset$, it thus follows that $b_i^{m.k}(h)(S_{-i}(h')) = 0$. Thus, b_i trivially satisfies Bayesian updating if the game moves from h to h'.

By combining the cases (i), (ii) and (iii), we see that b_i satisfies Bayesian updating.

Hence, for the strategy $s_i \in E_i^{m.k+1}$ there is some $b_i \in B_i$ that strongly believes $D_{-i}^{M.0}, D_{-i}^{M.1}, ..., D_{-i}^{m.k}$ such that s_i is optimal for b_i from period m onwards. That is, $s_i \in D_i^{m.k+1}$, which completes Case 1.

Case 2. Suppose that $k = 0$. If $m = M$, then $D^{m.k} = D^{M.0} = \times_{j \in I} S_j$. In that case, it would follow immediately that $s_i \in D_i^{m.k+1} = D_i^{M.1}$.

Suppose now that $k = 0$ and $m \leq M - 1$. Then, $D^{m.k} = D^{m.0} = D^{m+1.K_{m+1}}$. Since $s_i \in D_i^{m.0} = D^{m+1.K_{m+1}}$, there is some $b_i' \in B_i$ that strongly believes $D_{-i}^{M.0}, ..., D_{-i}^{m+1.K_{m+1}-1}$ such that s_i is optimal for b_i' from period $m+1$ onwards. Define the conditional belief vector b_i where

$$b_i(h) := \begin{cases} b_i^{m.0}(h), & \text{if } D_{-i}^{m.0} \cap S_{-i}(h) \neq \emptyset \\ b_i'(h), & \text{otherwise} \end{cases}$$

for every $h \in H_i$.

To show that $s_i \in D_i^{m.k+1} = D_i^{m.1}$, we show that $b_i \in B_i$, that b_i strongly believes $D_{-i}^{M.0}, ..., D_{-i}^{m+1.K_{m+1}-1}, D_{-i}^{m.0}$, and that s_i is optimal for b_i from period m onwards.

Note that, by construction, b_i strongly believes $D_{-i}^{M.0}, ..., D_{-i}^{m+1.K_{m+1}-1}$ and strongly believes $D_{-i}^{m.0}$. In the same way as for Case 1, it can be shown that b_i satisfies Bayesian updating, and hence $b_i \in B_i$.

We finally show that s_i is optimal for b_i from period m onwards. Take some information set $h \in H_i(s_i) \cap H^\tau$, for some $\tau \geq m$.

If $D_{-i}^{m.0} \cap S_{-i}(h) \neq \emptyset$, then $h \in H(D^{m.0}) \cap H_i(s_i)$ since $s_i \in D_i^{m.0}$. Thus, by construction, s_i is optimal for $b_i^{m.0}$ at h. Since $b_i^{m.0}(h) = b_i(h)$, we know that s_i is optimal for b_i at h.

If $D_{-i}^{m.0} \cap S_{-i}(h) = \emptyset$, then, by definition, $b_i(h) = b_i'(h)$. Since $D_{-i}^{m.0} \cap S_{-i}(h) = \emptyset$, it must necessarily be that $h \in H^\tau$ with $\tau \geq m+1$. Indeed, since $D_{-i}^{m.0}$ only imposes restrictions on actions at period $m+1$ and later, it follows that $D_{-i}^{m.0} \cap S_{-i}(h) \neq \emptyset$ for every $h \in H_i \cap H^m$. As s_i is optimal for b_i' from period $m+1$ onwards, we conclude that s_i is optimal for $b_i'(h)$ at h, and thus also for $b_i(h)$ at h.

Hence, s_i is optimal for b_i from period m onwards. Altogether, we see that $b_i \in B_i$, that b_i strongly believes $D_{-i}^{M.0}, ..., D_{-i}^{m+1.K_{m+1}-1}, D_{-i}^{m.0}$, and that s_i is optimal for b_i from period m onwards. Hence, $s_i \in D_i^{m.1} = D_i^{m.k+1}$, which completes Case 2.

By Cases 1 and 2 we thus conclude that $E_i^{m.k+1} \subseteq D_i^{m.k+1}$, and hence $E_i^{m.k+1} = D_i^{m.k+1}$. This completes the proof of the claim. $\Diamond$

Since, by construction, $sb(D^{m.k}) \subseteq E^{m.k+1}$, it follows by the claim that

$$sb(D^{m.k}) \subseteq E^{m.k+1} = D^{m.k+1} \subseteq D^{m.k}. \tag{10.6}$$

Consider now the case where $k = K_m$. Then the next step is $m - 1.0$ where, by definition, $D^{m-1.0} = D^{m.K_m}$. Hence, we trivially have that

$$sb(D^{m.K_m}) \subseteq D^{m-1.0} \subseteq D^{m.K_m}. \tag{10.7}$$

By (10.6) and (10.7) we conclude that $(D^{M.0}, ..., D^{1.K_1})$ is an elimination order for sb. This completes the proof. ■

10.5 Proofs of Section 8

Proof of Theorem 8.1. Let h^* be the information set that coincides with the initial history of Γ. Then, every strategy $\hat{s}_i \in S_i^{\hat{\Gamma}}(h^*)$ in the supergame that allows for h^* to be reached naturally induces a strategy $f_i(\hat{s}_i) \in S_i^{\Gamma}$ in the subgame, such that $\hat{s}_i$ and $f_i(\hat{s}_i)$ prescribe the same actions at all information sets in $H_i^{\hat{\Gamma}}(\hat{s}_i) \cap H_i^{\Gamma}$. Note that the mapping f_i is onto.

Now, consider a conditional belief vector $\hat{b}_i$ for player i in the supergame $\hat{\Gamma}$. Then, $\hat{b}_i$ naturally induces the conditional belief vector $g_i(\hat{b}_i)$ in the subgame Γ, where for every information set $h \in H_i^{\Gamma}$ and every opponents' strategy combination $(s_j)_{j \neq i} \in S_{-i}(h)$,

$$(g_i(\hat{b}_i))(h)((s_j)_{j \neq i}) := \sum_{(\hat{s}_j)_{j \neq i} \in S_{-i}^{\hat{\Gamma}}(h): f_j(\hat{s}_j) = s_j \text{ for all } j \neq i} \hat{b}_i(h)((\hat{s}_j)_{j \neq i}). \tag{10.8}$$

It may easily be verified that $g_i(\hat{b}_i)$ satisfies Bayesian updating whenever possible and that the mapping $g_i : B_i^{\hat{\Gamma}} \to B_i^{\Gamma}$ is onto.

Suppose, without loss of generality, that the supergame $\hat{\Gamma}$ starts at period 1, that the subgame Γ starts at the singleton information set h^* in period $L \geq 1$, and that the last period in which players are active is M in both $\hat{\Gamma}$ and Γ. It is of course possible that some terminal histories in $\hat{\Gamma}$ are longer than in Γ, but every terminal history in $\hat{\Gamma}$ that passes through information sets in Γ will be in Γ as well. This will therefore not affect $Z(S^{fbr}(\Gamma \mid \hat{\Gamma}))$.

For every player i, period $m \in \{1, ..., M\}$ and round k, let $\hat{B}_i^{m.k}$ and $\hat{S}_i^{m.k}$ be the set of conditional belief vectors and the set of strategies for player i that survive round k at period m in the backward and forward rationalizability procedure for $\hat{\Gamma}$. Similarly, we denote by $B_i^{m.k}$ and $S_i^{m.k}$ the set of conditional belief vectors and the set of strategies for player i that survive round k at period m in the backward and forward rationalizability procedure for Γ. For every period $m \in \{L, ..., M\}$, let K_m be the earliest round in this period at which both procedures terminate.

Claim. For every player i, period $m \in \{L, ..., M\}$ and round $k \in \{1, ..., K_m\}$, we have that

$$g_i(\hat{B}_i^{m.k}) = B_i^{m.k} \text{ and } f_i(\hat{S}_i^{m.k}) = S_i^{m.k}.$$

Proof of claim. By induction on $m.k$. We start by with $m.k = M.0$. Then, $\hat{B}_i^{M.0} = B_i^{\hat{\Gamma}}$, $B_i^{M.0} = B_i^{\Gamma}$, $\hat{S}_i^{M.0} = S_i^{\hat{\Gamma}}$ and $S_i^{M.0} = S_i^{\Gamma}$, which implies that

$$g_i(\hat{B}_i^{M.0}) = g_i(B_i^{\hat{\Gamma}}) = B_i^{\Gamma} = B_i^{M.0} \text{ and } f_i(\hat{S}_i^{M.0}) = f_i(S_i^{\hat{\Gamma}}) = S_i^{\Gamma} = S_i^{M.0},$$

since both g_i and f_i are onto.

Now, take some $m.k \neq M.0$, and assume that the claim holds for $m{+}1.K_{m+1}$ if $k = 0$ or for $m.k{-}1$ if $k \geq 1$. If $k = 0$ then, by definition, $\hat{B}_i^{m.0} = \hat{B}_i^{m+1.K_{m+1}}$, $B_i^{m.0} = B_i^{m+1.K_{m+1}}$, $\hat{S}_i^{m.0} = \hat{S}_i^{m+1.K_{m+1}}$ and $S_i^{m.0} = S_i^{m+1.K_{m+1}}$. Thus, the claim would hold by the induction assumption.

Assume now that $k \geq 1$ and that the claim holds for $m.k-1$. To show that $g_i(\hat{B}_i^{m.k}) \subseteq B_i^{m.k}$, take some $b_i \in g_i(\hat{B}_i^{m.k})$. Then, by definition, there is some $\hat{b}_i \in \hat{B}_i^{m.k}$ such that $b_i = g_i(\hat{b}_i)$. By definition of $\hat{B}_i^{m.k}$, it must be that $\hat{b}_i$ strongly believes $\hat{S}_{-i}^{m.k-1}$. In particular, for every $h \in H_i^\Gamma$ it must be that

$$\hat{b}_i(h)(\hat{S}_{-i}^{m.k-1}) = 1 \text{ whenever } \hat{S}_{-i}^{m.k-1} \cap S_{-i}^{\hat{\Gamma}}(h) \neq \emptyset. \tag{10.9}$$

Take such $h \in H_i^\Gamma$. We will show that

$$(g_i(\hat{b}_i))(h)(S_{-i}^{m.k-1}) = 1 \text{ whenever } S_{-i}^{m.k-1} \cap S_{-i}^{\Gamma}(h) \neq \emptyset. \tag{10.10}$$

Suppose that $S_{-i}^{m.k-1} \cap S_{-i}^{\Gamma}(h) \neq \emptyset$. By the induction assumption we know that $S_{-i}^{m.k-1} = f_{-i}(\hat{S}_{-i}^{m.k-1})$, where $f_{-i}(\hat{S}_{-i}^{m.k-1}) := \times_{j \neq i} f_j(\hat{S}_j^{m.k-1})$. Moreover, by construction, the sets $\hat{S}_{-i}^{m.k-1}$ and $f_{-i}(\hat{S}_{-i}^{m.k-1})$ induce the same actions in Γ. Since $m \geq L$, it follows that $\hat{S}_{-i}^{m.k-1} \cap S_{-i}^{\hat{\Gamma}}(h) \neq \emptyset$. By (10.9) we then conclude that $\hat{b}_i(h)(\hat{S}_{-i}^{m.k-1}) = 1$. It then follows by (10.8) that

$$(g_i(\hat{b}_i))(h)(f_{-i}(\hat{S}_{-i}^{m.k-1})) = 1.$$

Since, by the induction assumption, $S_{-i}^{m.k-1} = f_{-i}(\hat{S}_{-i}^{m.k-1})$, we conclude that

$$(g_i(\hat{b}_i))(h)(S_{-i}^{m.k-1}) = 1,$$

which establishes (10.10).

As this holds for every $h \in H_i^\Gamma$ with $S_{-i}^{m.k-1} \cap S_{-i}^{\Gamma}(h) \neq \emptyset$, it follows that $g_i(\hat{b}_i)$ strongly believes $S_{-i}^{m.k-1}$. Moreover, as $\hat{b}_i \in \hat{B}_i^{m.k-1}$ and, by the induction assumption, $g_i(\hat{B}_i^{m.k-1}) = B_i^{m.k-1}$, we conclude that $g_i(\hat{b}_i) \in B_i^{m.k-1}$. Hence, by definition, $b_i = g_i(\hat{b}_i) \in B_i^{m.k}$. Since this applies to every $b_i \in g_i(\hat{B}_i^{m.k})$, it follows that $g_i(\hat{B}_i^{m.k}) \subseteq B_i^{m.k}$.

To show that $B_i^{m.k} \subseteq g_i(\hat{B}_i^{m.k})$, take some $b_i \in B_i^{m.k}$. Then, in particular, $b_i \in B_i^{m.k-1}$. As, by the induction assumption, $B_i^{m.k-1} = g_i(\hat{B}_i^{m.k-1})$, there is some $\hat{b}_i \in \hat{B}_i^{m.k-1}$ such that $b_i = g_i(\hat{b}_i)$.

Moreover, as $b_i \in B_i^{m.k}$ we know, by definition, that b_i strongly believes $S_{-i}^{m.k-1}$. Hence, for every $h \in H_i^\Gamma$ we have that

$$b_i(h)(S_{-i}^{m.k-1}) = 1 \text{ whenever } S_{-i}^{m.k-1} \cap S_{-i}^{\Gamma}(h) \neq \emptyset.$$

As, by the induction assumption, $S_{-i}^{m.k-1} = f_{-i}(\hat{S}_{-i}^{m.k-1})$, it follows that

$$b_i(h)(f_{-i}(\hat{S}_{-i}^{m.k-1})) = 1 \text{ whenever } f_{-i}(\hat{S}_{-i}^{m.k-1}) \cap S_{-i}^{\Gamma}(h) \neq \emptyset. \tag{10.11}$$

By construction, $f_{-i}(\hat{S}_{-i}^{m.k-1})$ prescribes the same actions in the subgame Γ as $\hat{S}_{-i}^{m.k-1}$. Moreover, as $m \geq L$ and the first history of Γ is in period L, the set $\hat{S}_{-i}^{m.k-1}$ imposes no restrictions on actions before Γ starts. We therefore have that

$$f_{-i}(\hat{S}_{-i}^{m.k-1}) \cap S_{-i}^{\Gamma}(h) \neq \emptyset \text{ if and only if } \hat{S}_{-i}^{m.k-1} \cap S_{-i}^{\Gamma}(h) \neq \emptyset. \tag{10.12}$$

By combining (10.11) and (10.12), we see that for every $h \in H_i^\Gamma$,

$$b_i(h)(f_{-i}(\hat{S}_{-i}^{m.k-1})) = 1 \text{ whenever } \hat{S}_{-i}^{m.k-1} \cap S_{-i}^\Gamma(h) \neq \emptyset. \tag{10.13}$$

Since $b_i = g_i(\hat{b}_i)$ for some $\hat{b}_i \in \hat{B}_i^{m.k-1}$, it follows by (10.13) and (10.8) that we can choose $\hat{b}_i \in \hat{B}_i^{m.k-1}$ such that

$$\text{for every } h \in H_i^{\hat{\Gamma}}, \text{ we have } \hat{b}_i(h)(\hat{S}_{-i}^{m.k-1}) = 1 \text{ whenever } \hat{S}_{-i}^{m.k-1} \cap S_{-i}^\Gamma(h) \neq \emptyset.$$

This means that $\hat{b}_i$ strongly believes $\hat{S}_{-i}^{m.k-1}$. As $\hat{b}_i \in \hat{B}_i^{m.k-1}$, it follows that $\hat{b}_i \in \hat{B}_i^{m.k}$. Since $g_i(\hat{b}_i) = b_i$ we conclude that $b_i \in g_i(\hat{B}_i^{m.k})$. As this holds for every $b_i \in B_i^{m.k}$ we conclude that $B_i^{m.k} \subseteq g_i(\hat{B}_i^{m.k})$. Together with our insight above that $g_i(\hat{B}_i^{m.k}) \subseteq B_i^{m.k}$, it follows that $g_i(\hat{B}_i^{m.k}) = B_i^{m.k}$.

We next show that $f_i(\hat{S}_i^{m.k}) = S_i^{m.k}$ for every player i. To prove that $f_i(\hat{S}_i^{m.k}) \subseteq S_i^{m.k}$, take some $s_i \in f_i(\hat{S}_i^{m.k})$. Then, there is some $\hat{s}_i \in \hat{S}_i^{m.k}$ such that $s_i = f_i(\hat{s}_i)$. By definition of $\hat{S}_i^{m.k}$, there is some $\hat{b}_i \in \hat{B}_i^{m.k}$ such that $\hat{s}_i$ is optimal for $\hat{b}_i$ at every $h \in H_i^{\hat{\Gamma}}(\hat{s}_i)$ from period m onwards.

Take some $h \in H_i^\Gamma(s_i)$ from period m onwards. We will show that s_i is optimal for $(g_i(\hat{b}_i))(h)$ at h. Recall that, for every $\hat{s}_i' \in \hat{S}_i$, the transformed strategy $f_i(\hat{s}_i')$ induces the same actions as $\hat{s}_i$ at information sets in Γ. This insight, together with (10.8), leads to the conclusion that

$$u_i(\hat{s}_i', \hat{b}_i(h)) = u_i(f_i(\hat{s}_i'), (g_i(\hat{b}_i))(h) \text{ for all } \hat{s}_i' \in S_i^{\hat{\Gamma}}(h). \tag{10.14}$$

As $\hat{s}_i$ is optimal for $\hat{b}_i$ at h, we know that

$$u_i(\hat{s}_i, \hat{b}_i(h)) \geq u_i(\hat{s}_i', \hat{b}_i(h)) \text{ for all } \hat{s}_i' \in S_i^{\hat{\Gamma}}(h).$$

Together with (10.14), this yields

$$u_i(f_i(\hat{s}_i), (g_i(\hat{b}_i))(h) \geq u_i(f_i(\hat{s}_i'), (g_i(\hat{b}_i))(h) \text{ for all } \hat{s}_i' \in S_i^{\hat{\Gamma}}(h).$$

As $s_i = f_i(\hat{s}_i)$ and $f_i(S_i^{\hat{\Gamma}}(h)) = S_i^\Gamma(h)$, it follows that

$$u_i(s_i, (g_i(\hat{b}_i))(h) \geq u_i(s_i', (g_i(\hat{b}_i))(h) \text{ for all } s_i' \in S_i^\Gamma(h).$$

Since this holds for every $h \in H_i^\Gamma(s_i)$ from period m onwards, we conclude that s_i is optimal for $g_i(\hat{b}_i)$ at all $h \in H_i^\Gamma(s_i)$ from period m onwards.

As $\hat{b}_i \in \hat{B}_i^{m.k}$, we know from above that $g_i(\hat{b}_i) \in B_i^{m.k}$. Since s_i is optimal for $g_i(\hat{b}_i)$ at all $h \in H_i^\Gamma(s_i)$ from period m onwards, it follows, by definition, that $s_i = f_i(\hat{s}_i) \in S_i^{m.k}$. As this holds for every $s_i \in f_i(\hat{S}_i^{m.k})$, we conclude that $f_i(\hat{S}_i^{m.k}) \subseteq S_i^{m.k}$.

We next prove that $S_i^{m.k} \subseteq f_i(\hat{S}_i^{m.k})$. Take some $s_i \in S_i^{m.k}$. Then, by definition, there is some $b_i \in B_i^{m.k}$ such that s_i is optimal for b_i at every $h \in H_i^\Gamma(s_i)$ from period m onwards.

From above we know that $B_i^{m.k} = g_i(\hat{B}_i^{m.k})$, and hence there is some $\hat{b}_i \in \hat{B}_i^{m.k}$ with $b_i = g_i(\hat{b}_i)$. Choose some strategy $\hat{s}_i \in S_i^{\hat{\Gamma}}$ such that (i) $f_i(\hat{s}_i) = s_i$, and (ii) $\hat{s}_i$ is optimal for $\hat{b}_i$ at every $h \in H_i^{\hat{\Gamma}}(\hat{s}_i) \backslash H_i^{\Gamma}$ from period m onwards.

Now, take some $h \in H_i^{\Gamma}(s_i)$ from period m onwards. As s_i is optimal for b_i at h, we have that

$$u_i(s_i, b_i(h)) \geq u_i(s_i', b_i(h)) \text{ for all } s_i' \in S_i^{\Gamma}(h).$$

Since $s_i = f_i(\hat{s}_i)$ and $b_i = g_i(\hat{b}_i)$, we know that

$$u_i(f_i(\hat{s}_i), (g_i(\hat{b}_i))(h)) \geq u_i(f_i(\hat{s}_i'), (g_i(\hat{b}_i))(h)) \text{ for all } \hat{s}_i' \in S_i^{\hat{\Gamma}}(h).$$

Together with (10.14) we then conclude that

$$u_i(\hat{s}_i, \hat{b}_i(h)) \geq u_i(\hat{s}_i', \hat{b}_i(h)) \text{ for all } \hat{s}_i' \in S_i^{\hat{\Gamma}}(h),$$

which means that $\hat{s}_i$ is optimal for $\hat{b}_i$ at h. As this holds for every $h \in H_i^{\Gamma}(\hat{s}_i)$ from period m onwards, and since we know from above that $\hat{s}_i$ is optimal for $\hat{b}_i$ at every $h \in H_i^{\hat{\Gamma}}(\hat{s}_i) \backslash H_i^{\Gamma}$ from period m onwards, we conclude that $\hat{s}_i$ is optimal for $\hat{b}_i$ at every $h \in H_i^{\hat{\Gamma}}(\hat{s}_i)$ from period m onwards. This, together with the fact that $\hat{b}_i \in \hat{B}_i^{m.k}$, implies that $\hat{s}_i \in \hat{S}_i^{m.k}$.

As $s_i = f_i(\hat{s}_i)$, we conclude that $s_i \in f_i(\hat{S}_i^{m.k})$. Since this holds for every $s_i \in S_i^{m.k}$, we conclude that $S_i^{m.k} \subseteq f_i(\hat{S}_i^{m.k})$. Together with the insight above that $f_i(\hat{S}_i^{m.k}) \subseteq S_i^{m.k}$, it follows that $f_i(\hat{S}_i^{m.k}) = S_i^{m.k}$.

By induction on $m.k$, the proof of the claim is complete. $\diamondsuit$

To prove supergame monotonicity, take some $s_i \in S_i^{fbr}(\Gamma \mid \hat{\Gamma})$. Then, there is some conditional belief vector $\hat{b}_i \in \hat{B}_i^{1.K_1}$ that survives the forward and backward rationalizability procedure in $\hat{\Gamma}$ such that s_i is optimal for $\hat{b}_i$ at all $h \in H_i^{\Gamma}(s_i)$. In particular, we then know that $\hat{b}_i \in \hat{B}_i^{L.K_L}$, where L is the period where the subgame Γ starts. By the claim it follows that $g_i(\hat{b}_i) \in B_i^{L.K_L}$, which means that $g_i(\hat{b}_i) \in B_i^{fbr}(\Gamma)$.

Since s_i is optimal for $\hat{b}_i$ at all $h \in H_i^{\Gamma}(s_i)$ it follows by (10.14) that s_i is optimal for $g_i(\hat{b}_i)$ at all $h \in H_i^{\Gamma}(s_i)$. Since $g_i(\hat{b}_i) \in B_i^{fbr}(\Gamma)$ we conclude that $s_i \in S_i^{fbr}(\Gamma \mid \Gamma)$. As this holds for every $s_i \in S_i^{fbr}(\Gamma \mid \hat{\Gamma})$, we conclude that $S_i^{fbr}(\Gamma \mid \hat{\Gamma}) \subseteq S_i^{fbr}(\Gamma \mid \Gamma)$, and hence supergame monotonicity holds for the forward and backward rationalizability procedure. This completes the proof. ■

Proof of Corollary 8.1. Let $z \in Z(S^{fbr}(\Gamma \mid \hat{\Gamma}))$. Then, there is some $(s_i)_{i \in I} \in \times_{i \in I} S_i^{fbr}(\Gamma \mid \hat{\Gamma})$ such that z is induced by $(s_i)_{i \in I}$. By Theorem 8.1 we know that $(s_i)_{i \in I} \in \times_{i \in I} S_i^{fbr}(\Gamma \mid \Gamma)$, and hence $z \in Z(S^{fbr}(\Gamma \mid \Gamma))$ By applying Theorem 6.3 to Γ, we conclude that $z \in Z(S^{sr}(\Gamma \mid \Gamma))$. ■

Proof of Theorem 8.2. The proof for this result is very similar to the proof of Theorem 8.1, since both *backwards rationalizability* and *forward and backward rationalizability* proceed in a backward inductive fashion. The proof is therefore left to the reader. ■

Proof of Corollary 8.2. Proof is given in the text above the statement of this corollary. ■

Proof of Corollary 8.3. Follows immediately from Corollaries 8.1 and 8.2. ■

References

[1] Battigalli, P. (1997), On rationalizability in extensive games, *Journal of Economic Theory* **74,** 40–61.

[2] Battigalli, P. (2003), Rationalizability in infinite, dynamic games of incomplete information, *Research in Economics* **57,** 1–38.

[3] Battigalli, P. and N. de Vito (2021), Beliefs, plans, and perceived intentions in dynamic games, *Journal of Economic Theory* **195,** 105283.

[4] Battigalli, P. and M. Siniscalchi (1999), Hierarchies of conditional beliefs and interactive epistemology in dynamic games, *Journal of Economic Theory* **88,** 188–230.

[5] Battigalli, P. and M. Siniscalchi (2002), Strong belief and forward induction reasoning, *Journal of Economic Theory* **106,** 356–391.

[6] Battigalli, P. and M. Siniscalchi (2003), Rationalization and incomplete information, *B.E. Jounal of Theoretical Economics* **3**, 1–46.

[7] Catonini, E. (2019), Rationalizability and epistemic priority orderings, *Games and Economic Behavior* **114,** 101–117.

[8] Catonini, E. (2020), On non-monotonic strategic reasoning, *Games and Economic Behavior* **120,** 209–224.

[9] Chen, J. and S. Micali (2013), The order independence of iterated dominance in extensive games, *Theoretical Economics* **8,** 125–163.

[10] Cho, I.-K. (1987), A refinement of sequential equilibrium, *Econometrica* **55,** 1367–1389.

[11] Fukuda, S. (2023), The existence of universal qualitative belief spaces, Manuscript.

[12] Guarino, P. (2022), Topology-free type structures with conditioning events, Working paper.

[13] Heifetz, A., Meier, M. and B.C. Schipper (2013), Dynamic unawareness and rationalizable behavior, *Games and Economic Behavior* **81,** 50–68.

[14] Heifetz, A. and A. Perea (2015), On the outcome equivalence of backward induction and extensive form rationalizability, *International Journal of Game Theory* **44,** 37–59.

[15] Hillas, J. (1994), Sequential equilibria and stable sets of beliefs, *Journal of Economic Theory* **64,** 78–102.

[16] Kechris, A. (1995), *Classical Descriptive Set Theory,* Springer, Graduate Texts in Mathematics.

[17] Kohlberg, E. and J.-F. Mertens (1986), On the strategic stability of equilibria, *Econometrica* **54,** 1003–1037.

[18] Kreps, D.M. and R. Wilson (1982), Sequential equilibria, *Econometrica* **50**, 863–894.

[19] McLennan, A. (1985), Justifiable beliefs in sequential equilibria, *Econometrica* **53,** 889–904.

[20] Pearce, D.G. (1984), Rationalizable strategic behavior and the problem of perfection, *Econometrica* **52,** 1029–1050.

[21] Penta, A. (2015), Robust dynamic implementation, *Journal of Economic Theory* **160,** 280–316.

[22] Perea, A. (2012), *Epistemic Game Theory: Reasoning and Choice,* Cambridge University Press.

[23] Perea, A. (2014), Belief in the opponents' future rationality, *Games and Economic Behavior* **83,** 231–254.

[24] Perea, A. (2017), Order independence in dynamic games, *Epicenter Working Paper No. 8.*

[25] Perea, A. (2018), Why forward induction leads to the backward induction outcome: A new proof for Battigalli's theorem, *Games and Economic Behavior* **110,** 120–138.

[26] Reny, P. (1992), Backward induction, normal form perfection and explicable equilibria, *Econometrica* **60,** 627–649.

[27] Rubinstein, A. (1991), Comments on the interpretation of game theory, *Econometrica* **59**, 909–924.

[28] Selten, R. (1965), Spieltheoretische Behandlung eines Oligopolmodells mit Nachfrageträgheit, *Zeitschrift für die Gesammte Staatswissenschaft* **121,** 301–324, 667–689.

[29] Selten, R. (1975), Reexamination of the perfectness concept for equilibrium points in extensive games, *International Journal of Game Theory* **4,** 25–55.

[30] Shimoji, M. and J. Watson (1998), Conditional dominance, rationalizability, and game forms, *Journal of Economic Theory* **83,** 161–195.

[31] van Damme, E. (1984), A relation between perfect equilibria in extensive form games and proper equilibria in normal form games, *International Journal of Game Theory* **13**, 1–13.

[32] von Neumann, J. and O. Morgenstern (1953), *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, NJ.

[33] Zermelo, E. (1913), Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels, *Proceedings Fifth International Congress of Mathematicians* **2,** 501–504.