International Journal of Game Theory
https://doi.org/10.1007/s00182-020-00733-1

ORIGINAL PAPER

Should I remember more than you? Best responses to factored strategies

René Levínský[1] · Abraham Neyman[2] · Miroslav Zelený[3]

Accepted: 26 August 2020

Abstract

In this paper we offer a new, unifying approach to modeling strategies of bounded complexity. In our model, the strategy of a player in a game does not directly map the set H of histories to the set of her actions. Instead, the player's perception of H is represented by a map $\varphi\colon H \to X$, where X reflects the "cognitive complexity" of the player, and the strategy chooses its mixed action at history h as a function of $\varphi(h)$. In this case we say that φ is a factor of a strategy and that the strategy is φ-factored. Stationary strategies, strategies played by finite automata, and strategies with bounded recall are the most prominent examples of factored strategies in multistage games.A factor φ is recursive if its value at history h' that follows history h is a function of $\varphi(h)$ and the incremental information $h' \setminus h$. For example, in a repeated game with perfect monitoring, a factor φ is recursive if its value $\varphi(a_1, \ldots, a_t)$ on a finite string of action profiles $(a_1, \ldots, a_t)$ is a function of $\varphi(a_1, \ldots, a_{t-1})$ and a_t.We prove that in a discounted infinitely repeated game and (more generally) in a stochastic game with finitely many actions and perfect monitoring, if the factor φ is recursive, then for every profile of φ-factored strategies there is a pure φ-factored strategy that is a best reply, and if the stochastic game has finitely many states and actions and the factor φ has a finite range then there is a pure φ-factored strategy that is a best reply in all the discounted games with a sufficiently large discount factor.

Keywords Bounded rationality · Factored strategies · Bounded recall strategies · Finite automata

René Levínský: Research was supported by GAČR Grant 17-19672S and by the Ministry of Education, Youth and Sports of the Czech Republic through the project $SHARE - CZ + (CZ.02.1.01/0.0/0.0/16_0 13/0001740)$. Abraham Neyman: Research was supported in part by Israel Science Foundation grant 1123/06. Miroslav Zelený: Research was supported by MSM research project 0021620839 financed by MSMT and GAČR Grant 17-19672S.

Extended author information available on the last page of the article

1 Introduction

There are two widely studied approaches to modeling strategies of bounded complexity in repeated games. One approach is to assume that players have stationary bounded recall Aumann (1981), Lehrer (1988), Aumann and Sorin (1989). Such players have imperfect consciousness of the actual stage of the game, and their actions in the current stage game rely only on the signals, e.g., players' actions, of the last k rounds. The strategies of such players are called stationary bounded recall strategies, henceforth, SBR strategies. The other approach is to assume that players have stationary finite memory Neyman (1985), Rubinstein (1986), Abreu and Rubinstein (1988), Ben-Porath (1993). Such players track a finite summary of the history of play and update their summaries as a function of the current memory and the new input. The strategies of such players are implementable by finite automata (Moore machines).

Both approaches provide a measure of the complexity of the strategy. Under the bounded recall approach, the complexity of a strategy is described by the "depth of recall" and under the finite memory approach the complexity of a strategy is measured by the minimal number of states the automaton must have to play the given strategy.

In this paper, the question raised by Kalai (1990), "What information system (size and structure) should a player maintain when playing a strategic game?," is examined in the context of strategies of bounded complexity. Explicitly, we study the minimal complexity of a strategy that is the best response to a strategy of a given complexity. Abreu and Rubinstein (1988) shows that for every finite automaton A_1 of Player 1, there exists a finite automaton A_2 of Player 2 whose number of states is less than or equal to the number of states of A_1, such that A_2 maximizes its own payoff in the discounted repeated game against A_1. Here, we address this question in the broader context of the newly defined concept of factored strategies.

In addition, we study this question not only in the classic repeated game context, namely, in supergames with finitely many stage actions and perfect monitoring, but also in the stochastic game context.

Let us also remark that relationships between the minimal size of an automaton that implements a strategy and the complexity of the strategy were investigated by Kalai and Stanford (1988) who studied the complexity of strategies forming equilibria.

A strategy is any map from the player's information sets to the set of feasible actions. In the repeated game model with perfect monitoring the information sets are the possible histories, i.e., the finite strings of action profiles.

Under our bounded rationality approach, the player may not be cognitively capable of distinguishing two distinct histories. Hence, the player can base her actions only on some equivalence classes of indistinguishable histories.[1]

Let φ be the map that assigns to any possible history h its equivalence class of indistinguishable histories x. Hence, φ is a function from the set H of all possible histories onto the set X of equivalence classes of indistinguishable histories. Naturally, we are interested in cases where the function φ is not injective. A strategy whose action at any history h is a function of $\varphi(h)$ is called a *φ-factored strategy*.

[1] Partitions of the space of histories have already been used for other purposes in the theory of Markovian strategies Maskin and Tirole (2001).

Journal: 182 Article No.: 0733 TYPESET DISK LE CP Disp.:2020/10/1 Pages: 20 Layout: Small-Ex

Examples of factored strategies in the repeated game are SBR strategies and strategies that are implementable by an automaton. For example, let us illustrate that a stationary bounded k-recall strategy is a factored strategy: it does not distinguish between two different histories h and h' that are identical in the last k coordinates; hence, its choice of an action depends on the image of the factor φ that maps a history to its last k action profiles.

The question that arises is, what are the game models and the factors φ such that for any φ-factored strategy of Player 1, Player 2 has a best reply that is a φ-factored strategy?

We show an example of a discounted repeated game, a factor φ, and a profile of φ-factored strategies for which no φ-factored strategy is a best reply. Hence, even in the discounted repeated game model an additional assumption on the factor φ is needed in order to derive a positive result.

The additional assumption is recursiveness. A factor in the repeated game is *recursive* if its value on any finite string of actions $(a_1, \ldots, a_t)$ is a function of a_t and its value on $(a_1, \ldots, a_{t-1})$. Note that the factor of a SBR strategy and of a strategy that is implementable by an automaton (with deterministic transitions) is recursive.

Our main result asserts that if the factor φ is recursive, then, for any profile of φ-factored strategies in the discounted repeated game with finitely many stage actions, there is a best reply that is a pure φ-factored strategy. This result holds not only in the classic repeated game context but also in the stochastic game context.

Our main result generalizes analogous results of Blackwell (1962) and Derman (1965) for a Markov decision process (MDP), and the proof relies on the same ideas as those used in the proofs of Blackwell (1962) and Derman (1965).

All relevant notions will be defined and discussed in the next section. Section 3 introduces the concept of factored strategies and presents examples. Section 4 contains the main results and their proofs. In Sect. 5 we connect our results with well-known theorems of Derman and Blackwell, Sect. 6 concludes.

2 The game models

If X is a finite or countable set (or a measurable space), then $\Delta(X)$ denotes the set of all probabilities on X. For ease of exposition we focus on multistage two-person games with perfect monitoring. The results of the present paper on the best reply can be extended (from two-person games) to n-person games ($n > 2$) by considering Players $1, 2, \ldots, n-1$ as a single player.

2.1 Supergames

We start by recalling the model of the two-person supergame with finite action sets. Let $G = \langle A_1, A_2, u_1, u_2 \rangle$ be a strategic game, where A_i is the nonempty finite set of actions of player $i \in \{1, 2\}$ and $u_i \colon A_1 \times A_2 \to \mathbb{R}$ is the payoff function of player i. The corresponding supergame G^∞ is played as follows. In each period $t \in \mathbb{N} = \{1, 2, 3, \ldots\}$, Players 1 and 2 make simultaneous and independent moves

Journal: 182 Article No.: 0733 TYPESET DISK LE CP Disp.:2020/10/1 Pages: 20 Layout: Small-Ex

$a_t^i \in A_i$, $i = 1, 2$. A *play* of the supergame is a sequence of action profiles $(a_t)_{t=1}^{\infty}$ with $a_t = (a_t^1, a_t^2) \in A := A_1 \times A_2$. A play $(a_t)_{t=1}^{\infty}$ defines a stream $(u_i(a_t))_{t=1}^{\infty}$ of payoffs to player i.

We assume perfect monitoring. Hence, a *pure strategy for player i in the supergame* G^{∞} is a mapping $\sigma\colon A^* \to A_i$, where A^* denotes the set of all finite sequences of elements from A (including the empty one). A pure strategy σ plays in the t-th round the action $\sigma(a_1, \ldots, a_{t-1})$, where $(a_1, \ldots, a_{t-1}) \in A^{t-1}$ is the sequence of actions that have already been played.

A *behavioral strategy for player i in the supergame* G^{∞} is a mapping $\sigma\colon A^* \to \Delta(A_i)$. A behavioral strategy σ plays in the t-th round an action $a_t^i \in A_i$ with the probability $\sigma(a_1, \ldots, a_{t-1})(a_t^i)$, where $(a_1, \ldots, a_{t-1}) \in A^{t-1}$ is the sequence of actions that have already been played. Pure strategies can be viewed as a special case of behavioral strategies by identifying A_i with the Dirac measures on A_i. This point of view will be used throughout the paper.

2.2 Supergames with time-dependent stage games

The previous concept can be generalized as follows. Let $\Gamma = \{\langle A_1(t), A_2(t), u_1(t), u_2(t)\rangle\}$ be a sequence of stage games. The game Γ is played as follows. In each period $t \in \mathbb{N}$, Players 1 and 2 make simultaneous and independent moves $a_t^i \in A_i(t)$, $i = 1, 2$. A *play* of Γ is a sequence of action profiles $(a_t)_{t=1}^{\infty}$ with $a_t = (a_t^1, a_t^2) \in A(t) := A_1(t) \times A_2(t)$. A play $(a_t)_{t=1}^{\infty}$ defines a stream $(u_i(t)(a_t))_{t=1}^{\infty}$ of payoffs to player i. The pure and behavioral strategies of player i in Γ are defined as a family of mappings σ_t, $t \geq 1$, from $\times_{s<t} A(s)$ to $\Delta(A_i(t))$.

2.3 Stochastic games

A two-person *stochastic game* with finite action sets is a 5-tuple $\Gamma = \langle S, A, u, p, \mu\rangle$, where

- S is a nonempty set, called the *state space*;
- $A(z) = A_1(z) \times A_2(z)$ is a cartesian product of two nonempty finite sets. It is the set of *action pairs*: for every state $z \in S$, $A_i(z)$ is the finite set of feasible actions of player $i \in \{1, 2\}$ in state z;
- $u = (u_1, u_2)$ is a function from the set of pairs (z, a) with $z \in S$ and $a \in A(z)$ to $\mathbb{R}^2$. It is the *payoff function*: $u_i(z, a)$ is the payoff to player i when the action pair $a \in A(z)$ is played in state $z \in S$;
- p is a *transition function*: for each state $z \in S$ and each action pair $a \in A(z)$, $p(\cdot \mid z, a) \in \Delta(S)$ is the distribution of the next state; i.e., $p(z' \mid z, a)$ is the probability of moving to state z' if the players played the action pair a in state z;
- $\mu \in \Delta(S)$ is the distribution of the initial state.

A *play* of the stochastic game Γ is a sequence of states and action pairs $(z_1, a_1, \ldots, z_t, a_t, \ldots)$ with $a_t \in A(z_t)$. The set of all plays in Γ is denoted by H^{∞}, i.e.,

$$H^{\infty} = \big\{(z_1, a_1, \ldots, z_t, a_t, \ldots) : z_t \in S \text{ and } a_t \in A(z_t) \text{ for every } t \in \mathbb{N}\big\}.$$

Additional measurability conditions, which are not stated here, are assumed in the case where the state space is a measurable space.

Consider a two-person stochastic game with perfect monitoring. In such a game, each player observes the sequence $(z_1, a_1, \ldots, a_{t-1}, z_t)$ of past states and action pairs before choosing her action in state t. Hence, a *pure strategy of player i* specifies her action $a_t^i \in A_i(z_t)$ as a function of the sequence $(z_1, a_1, \ldots, a_{t-1}, z_t)$ of past states and action pairs. Similarly, a *behavioral strategy of player i* specifies the probability of her action $a_t^i \in A_i(z_t)$ as a function of the sequence $(z_1, a_1, \ldots, a_{t-1}, z_t)$ of past states and action pairs.

A pair σ^1 and σ^2 of behavioral strategies of Players 1 and 2, respectively, defines a probability distribution P_{σ^1,σ^2} on the space of plays H^∞ as follows. The distribution of z_1 is μ. The conditional probability that $a_t = (a^1, a^2) \in A(z_t)$, given $z_1, a_1, \ldots, z_t$, is

$$\sigma^1(z_1, a_1, \ldots, z_t)\left(a^1\right) \cdot \sigma^2(z_1, a_1, \ldots, z_t)\left(a^2\right).$$

The conditional distribution of z_{t+1}, given $z_1, a_1, \ldots, z_t, a_t$, is $p(\cdot \mid z_t, a_t)$.

Similarly, a finite initial segment h of a play of the stochastic game along a pair σ^1 and σ^2 of behavioral strategies (of Players 1 and 2, respectively) defines a probability distribution $P_{(\sigma^1,\sigma^2|h)}$ on the set $H(h)$ of all plays that extend the finite sequence h. The expectation with respect to the probability distribution $P_{(\sigma^1,\sigma^2|h)}$ is denoted by $E_{(\sigma^1,\sigma^2|h)}$.

2.3.1 Payoffs

Let σ^1 and σ^2 be strategies of Players 1 and 2, respectively, in the stochastic game Γ and let $h = (z_1, a_1, \ldots, z_\ell)$, or $h = (z_1, a_1, \ldots, a_{\ell-1})$, be a finite history. The subgame that follows the initial history h is denoted by $(\Gamma \mid h)$.

The β *-discounted payoff to player i*, $0 \le \beta < 1$, in $(\Gamma \mid h)$ is defined by

$$V_\beta^i(\sigma^1, \sigma^2 \mid h) = E_{(\sigma^1,\sigma^2|h)}\left(\sum_{t=l}^{\infty} \beta^{t-l} u_i(z_t, a_t)\right).$$

2.3.2 Subclasses of stochastic games

A *Markov decision process* (henceforth *MDP*) is a one-person stochastic game; hence, it is a special case of a stochastic game.

A supergame is a stochastic game with a single state; hence, it is a special case of a stochastic game.

Similarly, a supergame with time-dependent stage games can be viewed as a stochastic game whose state space is the set of positive integers $\mathbb{N}$ and the transitions are deterministic: $\mathbb{N} \ni t \mapsto t + 1$.

Therefore, the β-discounted payoff is well defined for a supergame, as well as for a supergame with time-dependent stage games as long as the stage payoffs are either bounded or grow at a subexponential rate in t.

Journal: 182 Article No.: 0733 TYPESET DISK LE CP Disp.:2020/10/1 Pages: 20 Layout: Small-Ex

Therefore, results for stochastic games with finitely many states have direct consequences for supergames and results for stochastic games with countably many states have direct consequences for supergames with time-dependent stage games.

3 Factored strategies

In this section we define the concept of a factored strategy. This concept plays a central role in the results of the paper. For simplicity, we start with the introduction of the concept in the supergame model, and thereafter continue with the definition in the more general model of a stochastic game.

3.1 Factored strategies in supergames

Let F denote the set of all information sets. In a supergame with perfect monitoring F is the set of all finite histories A^*. Let X be a set and let φ be a mapping from F to X. We say that a behavioral strategy σ is a φ- *factored* strategy for player i in the supergame G^∞ if there is an *action function* $\omega\colon X \to \Delta(A_i)$ such that $\sigma = \omega \circ \varphi$.

The factor φ is called *recursive* if there is a function $g\colon X \times A \to X$ such that $\varphi(a_1, \ldots, a_t) = g(\varphi(a_1, \ldots, a_{t-1}), a_t)$. Roughly speaking, if a player's cognition is captured by a recursive factor then she never learns something she previously ignored.

We continue by recalling the concepts of time-independent bounded recall strategies and of (their superset of) strategies that are implementable by time-independent automata with finitely many states. Such strategies are examples of factored strategies with a recursive factor.

Let $k \in \mathbb{N}$. A *behavioral k-SBR strategy for player i* in the supergame G^∞ is defined by a pair (e, ω), where $e = (e_1, e_2, \ldots, e_k) \in A^k$ and a mapping $\omega\colon A^k \to \Delta(A_i)$. The strategy that is defined by (e, ω) is played as follows. If moves $a_1, \ldots, a_l \in A$ have been played, then player i takes the sequence b of the last k elements of the sequence $(e_1, \ldots, e_k, a_1, \ldots, a_l)$, and his $(l+1)$-th move is $a \in A_i$ with the (conditional) probability $\omega(b)(a)$. A *pure k-SBR strategy for player i* in the supergame G^∞ is defined in a similar way.

Defining $\varphi(a_1, \ldots, a_l)$ by the last k elements of the finite sequence of action profiles

$$(e_1, \ldots, e_k, a_1, \ldots, a_l),$$

we see that the k-SBR strategy σ defined above obeys $\sigma = \omega \circ \varphi$, and the factor φ is recursive; thus, σ is a factored strategy with a recursive factor φ that has a finite range.

We say that a (pure) behavioral strategy σ is a *(pure) behavioral SBR strategy* if σ is a (pure) behavioral k-SBR strategy for some $k \in \mathbb{N}$.

A *behavioral automaton* (for Player 1 in the supergame G^∞) is a quadruple $\langle M, m^*, \alpha, \tau \rangle$, where M is a nonempty set (the state space), $m^* \in M$ is the initial state, $\alpha\colon M \to \Delta(A_1)$ is a probabilistic action function, and $\tau\colon M \times A \to M$ is a transition function. A *k-state behavioral automaton* is a behavioral automa-

Journal: 182 Article No.: 0733 TYPESET DISK LE CP Disp.:2020/10/1 Pages: 20 Layout: Small-Ex

ton where the set M has k elements. A behavioral automaton $\langle M, m^*, \alpha, \tau\rangle$ defines a behavioral strategy σ^1 (for Player 1) inductively: $m_1 = m^*$, $\sigma^1(\emptyset) = \alpha(m_1)$, $\sigma^1(a_1, \ldots, a_{t-1}) = \alpha(m_t)$, where $m_t = \tau(m_{t-1}, a_{t-1})$. A behavioral automaton $\langle M, m^*, \alpha, \tau\rangle$ defines a factored strategy with a recursive factor φ, where $X = M$, $\varphi(\emptyset) = m^*$, $\varphi(a_1, \ldots, a_t) = \tau(\varphi(a_1, \ldots, a_{t-1}), a_t)$, and $\omega = \alpha$. A *k-state (deterministic) automaton* is defined by replacing $\Delta(A_1)$ (in the definition of α) by A_1.

3.2 Factored strategies in stochastic games

A factor φ in a stochastic game keeps track of the current state as well as of some statistics of the past history. That is, φ maps the set $F := S \times (A \times S)^*$ of all information sets in the stochastic game to a set X and the map φ obeys $\varphi(z_1, a_1, \ldots, z_t) \neq \varphi(z'_1, a'_1, \ldots, z'_{t'})$ whenever $z_t \neq z'_{t'}$. Equivalently, X is the disjoint union of sets X_z and φ is a mapping from F to X such that $\varphi(z_1, a_1, \ldots, z_t) \in X_{z_t}$.

A behavioral strategy σ is a *φ-factored* strategy of player i in the stochastic game if there is an *action function* $\omega\colon X \to \Delta(A_i)$ such that

$$\sigma(z_1, a_1, \ldots, a_{t-1}, z_t) = \omega\big(\varphi(z_1, a_1, \ldots, z_{t-1}, a_{t-1}, z_t)\big).$$

In a stochastic game, a factor φ is called *recursive* if there is a function $g\colon X \times A \times S \to X$ such that

$$\varphi(z_1, a_1, \ldots, z_t, a_t, z_{t+1}) = g(\varphi(z_1, a_1, \ldots, z_{t-1}, a_{t-1}, z_t), a_t, z_{t+1}).$$

3.3 Time-dependent factored strategies

A time-dependent factor φ in a multistage game keeps track of the stage number as well as of some statistic of the past history. That is, if h and h' are two information sets at two distinct stages t and t', then $\varphi(h) \neq \varphi(h')$. Hence, a time-dependent factor φ is a factor φ that maps the set of histories of a multistage game into the disjoint union $X = \bigcup_{t\in\mathbb{N}} X_t$, where a history up to the play of the t-th stage is mapped into X_t.

A *time-dependent factored strategy* for player i is a factored strategy with respect to a time-dependent factor φ.

We continue by recalling the concepts of a time-dependent action automaton, a time-dependent transition automaton, and a time-dependent (action and transition) automaton. A strategy that is defined by a time-dependent automaton is a time-dependent factored strategy.

A *time-dependent action automaton* is defined by replacing the action function α by a sequence of action functions α_t, $t \geq 1$, where α_t defines the action at stage t [see, e.g.,(Neyman 1997)]. Similarly, a *time-dependent transition automaton* is obtained by replacing the (stationary) transition function τ with a sequence of time-dependent transitions τ_t, $t \geq 1$, where τ_t defines the transition at stage t. Finally, a *time-dependent (action and transition) automaton* in the supergame G^∞ is a quadruple $\langle M, m^*, (\alpha_t)_{t=1}^\infty, (\tau_t)_{t=1}^\infty\rangle$, where M is a nonempty set (the state space), $m^* \in M$ is the

Journal: 182 Article No.: 0733 TYPESET DISK LE CP Disp.:2020/10/1 Pages: 20 Layout: Small-Ex

initial state, $\alpha_t : M \to \Delta(A_1)$ is a probabilistic action function, and $\tau_t : M \times A \to M$ is a deterministic transition function.

A time-dependent automaton $\langle M, m^*, (\alpha_t)_{t=1}^{\infty}, (\tau_t)_{t=1}^{\infty} \rangle$ defines a behavioral strategy σ^1 (for Player 1) inductively: $m_1 = m^*, \sigma^1(\emptyset) = \alpha_1(m_1)$, and $\sigma^1(a_1, \ldots, a_{t-1}) = \alpha_t(m_t)$, where $m_t = \tau_t(m_{t-1}, a_{t-1})$.

Note that a time-dependent automaton $\langle M, m^*, (\alpha_t)_{t=1}^{\infty}, (\tau_t)_{t=1}^{\infty} \rangle$ defines the same strategy as the automaton $\langle M \times \mathbb{N}, m^{**}, \alpha, \tau \rangle$ with $m^{**} = (m^*, 1)$, $\alpha(m, t) = \alpha_t(m)$, and $\tau((m, t), a) = (\tau_t(m, a), t + 1)$. Therefore, the corresponding strategy is a recursive φ-factored strategy, where $\varphi : A^* \to M \times \mathbb{N}$ is given by $\varphi(a_1, \ldots, a_t) = \tau(\varphi(a_1, \ldots, a_{t-1}), a_t)$ and $\omega = \alpha$.

We recall the concept of a strategy with time-dependent recall. Let $k : \mathbb{N} \to \mathbb{N}$ be a function with $k(t) < t$ for every $t \in \mathbb{N}$. A *behavioral* (respectively, *pure*) *k -SBR strategy* in the supergame G^{∞} is defined analogously to the above case but the action at stage t depends only on the last $k(t)$ stage actions [see, e.g., (Neyman and Okada 2009)]. Let σ be such a strategy. Setting $\varphi(a_1, \ldots, a_t) = (a_{t-k(t)+1}, \ldots, a_t)$, we easily see that σ is φ-factored.

We now define the concept of a time-dependent strategy with time-dependent recall. Let $k : \mathbb{N} \to \mathbb{N}$ be a function with $k(t) < t$ for every $t \in \mathbb{N}$. A *behavioral* (respectively, *pure*) *time-dependent k -BR strategy* in the supergame G^{∞} is defined analogously to the above case but the action at stage t depends only on t and the last $k(t)$ stage actions. Let σ be such a strategy. Setting $\varphi(a_1, \ldots, a_t) = \big(t, (a_{t-k(t)+1}, \ldots, a_t)\big)$, we easily see that σ is φ-factored. Moreover, if $k(t + 1) \leq k(t) + 1$ for every $t \in \mathbb{N}$, then φ is recursive.

4 The Main results

In this section we state and prove the main results in the stochastic game context. In a later section we will spell out the implications of the main results in the special cases of a supergame, a supergame with time-dependent stage games, and an MDP.

It should be noted that the main result is analogous to, and implies, known results in the theory of Markov decision processes, and the proof follows the same ideas as the proofs of the analogous result for an MDP. However, there is a slight difference, on which we will comment in Remark 5.4.

The first result, Theorem 4.1, gives conditions on a stochastic game such that for every recursive factor φ, a φ-factored strategy of Player 1, and a discount factor β, Player 2 has a φ-factored strategy that is a best reply in the β-discounted game.

Theorem 4.1 *Let $\Gamma = \langle S, A, u, p, \mu \rangle$ be a two-person stochastic game with countably many states,[2] finitely many actions in each state, and a bounded payoff function u_2. Let σ^1 be a φ-factored behavioral strategy of Player 1 in Γ.*

If φ is recursive, then, for every $\beta \in (0, 1)$, there exists a φ-factored pure strategy σ^2 of Player 2, such that for every behavioral strategy ρ of Player 2 in Γ and every

[2] The assumption of countably many states is not essential here. The result holds also for stochastic games with a continuum of states and some measurability assumptions. For simplicity, and in particular to avoid the need to spell out the needed measurability assumptions, we assume countably many states.

Journal: 182 Article No.: 0733 TYPESET DISK LE CP Disp.:2020/10/1 Pages: 20 Layout: Small-Ex

initial history $h = (z_1, a_1, \ldots, z_l)$ *we have*

$$V_\beta^2\left(\sigma^1,\ \sigma^2 \mid h\right) \geq V_\beta^2\left(\sigma^1, \rho \mid h\right).$$

The second result, Theorem 4.2, gives conditions on a stochastic game such that for every recursive factor φ and a φ-factored strategy of Player 1, Player 2 has a φ-factored strategy that is a best reply in each one of the β-discounted games with a sufficiently large discount factor $\beta < 1$.

Theorem 4.2 *Let* $\Gamma = \langle S, A, u, p, \mu\rangle$ *be a two-person stochastic game with finitely many actions in each state. Let* φ *be a recursive factor with a finite range*[3] *and let* σ^1 *be a* φ*-factored behavioral strategy of Player 1 in* Γ.

Then, there is a φ*-factored pure strategy* σ^2 *of Player 2 and a discount factor* $\beta_0 \in (0, 1)$*, such that for every behavioral strategy* ρ *(of Player 2 in* Γ*) and every initial history* $h = (z_1, a_1, \ldots, z_l)$ *we have, for every* $\beta \in [\beta_0, 1)$*,*

$$V_\beta^2\left(\sigma^1, \sigma^2 \mid h\right) \geq V_\beta^2\left(\sigma^1, \rho \mid h\right).$$

The next result, Theorem 4.3, gives conditions on a stochastic game such that for every recursive factor φ and a φ-factored strategy of Player 1, Player 2 has a φ-factored strategy that is (a) a best reply in each one of the β-discounted games with a sufficiently large discount factor $\beta < 1$ as well as in the limiting-average payoff game, and (b) an approximate (within a possible error $\varepsilon > 0$) best reply in each of the sufficiently long finite-horizon stochastic games.

Theorem 4.3 *Let* $\Gamma = \langle S, A, u, p, \mu\rangle$ *be a two-person stochastic game with finitely many actions in each state. Let* φ *be a recursive factor with a finite range and let* σ^1 *be a* φ*-factored behavioral strategy of Player 1 in* Γ.

Let σ^2 *be a* φ*-factored strategy and* $\beta_0 \in (0, 1)$*, such that for every behavioral strategy* ρ *(of Player 2 in* Γ *) and every finite history* h*,* $V_\beta^2\left(\sigma^1, \sigma^2 \mid h\right) \geq V_\beta^2\left(\sigma^1, \rho \mid h\right)$ *for every* $\beta \in [\beta_0, 1)$*.*[4] *Then, there is a function* $N\colon (0, \infty) \to \mathbb{N}$*, such that for every strategy* ρ *of Player 2 in* Γ *and every initial history* $h = (z_1, a_1, \ldots, z_l)$ *we have*

$$E_{(\sigma^1,\sigma^2|h)}\left(\liminf_{n\to\infty} \frac{1}{n}\sum_{t=l}^{l+n-1} u_2(z_t, a_t)\right) \geq E_{(\sigma^1,\rho|h)}\left(\limsup_{n\to\infty} \frac{1}{n}\sum_{t=l}^{l+n-1} u_2(z_t, a_t)\right), \tag{1}$$

and

$$E_{(\sigma^1,\sigma^2|h)}\left(\frac{1}{n}\sum_{t=l}^{l+n-1} u_2(z_t, a_t)\right) \geq E_{(\sigma^1,\rho|h)}\left(\frac{1}{n}\sum_{t=l}^{l+n-1} u_2(z_t, a_t)\right) - \varepsilon \quad \forall n \geq N(\varepsilon). \tag{2}$$

[3] The existence of a factor with a finite range implies that the stochastic game has finitely many states.

[4] The existence of such a strategy σ^2 is guaranteed by Theorem 4.2.

Journal: 182 Article No.: 0733 TYPESET DISK LE CP Disp.:2020/10/1 Pages: 20 Layout: Small-Ex

Proof of Theorem 4.1

Fix $\beta \in (0, 1)$. Let σ^1 be a strategy of Player 1 in Γ and let $h = (z_1, a_1, \ldots, z_t)$ be a finite history. Define $V_\beta^2(\sigma^1 \mid h)$ by

$$V_\beta^2\left(\sigma^1 \mid h\right) = \sup_{\sigma^2} V_\beta^2\left(\sigma^1, \sigma^2 \mid h\right),$$

where the supremum is taken over all strategies of Player 2. Note that a strategy τ of Player 2 is a best reply to σ^1 in the β-discounted game $(\Gamma \mid h')$ whenever $V_\beta^2(\sigma^1, \tau|h') = V_\beta^2(\sigma^1 \mid h')$.

Assume that τ is a strategy such that for every finite history $h = (z_1, a_1, \ldots, z_t)$,

$$E_{(\sigma^1, \tau|h)}\left(u_2(z_t, a_t) + \beta V_\beta^2\left(\sigma^1 \mid h, a_t, z_{t+1}\right)\right) = V_\beta^2\left(\sigma^1 \mid h\right). \tag{3}$$

Fix a finite history $h' = (z_1, \ldots, a_{t'-1}, z_{t'})$ of the stochastic game. For $t \geq t'$ we denote by U_t the expectation w.r.t. $P_{(\sigma^1, \tau|h')}$ of $V_\beta^2(\sigma^1 \mid h)$ where h is the finite history $h = (h', \ldots, z_t)$ up to the play at stage t. By taking the expectation w.r.t. $P_{(\sigma^1, \tau|h')}$ in equality (3), we deduce that

$$E_{(\sigma^1, \tau|h')} u_2(z_t, a_t) = U_t - \beta U_{t+1}.$$

Multiplying this inequality by $\beta^{t-t'}$ and summing over all $t \geq t'$, we conclude that

$$\begin{aligned} E_{(\sigma^1, \tau|h')}\Big(\sum_{t=t'}^{\infty} \beta^{t-t'} u_2(z_t, a_t)\Big) &= \sum_{t=t'}^{\infty} \beta^{t-t'} U_t - \sum_{t=t'+1}^{\infty} \beta^{t-t'} U_t \\ &= U_{t'} = V_\beta^2\left(\sigma^1 \mid h'\right). \end{aligned}$$

Hence, τ is a best reply to σ^1 in the β-discounted game $(\Gamma \mid h)$.

Let φ be a recursive factor and σ^1 be a φ-factored strategy for Player 1 in the stochastic game Γ. Thus, φ maps the histories $(z_1, \ldots, z_t)$ to a set X, and there exist functions $\omega\colon X \to \Delta(A_1)$ and $g\colon X \times A \times S \to X$ such that

$$\sigma^1(z_1, a_1, \ldots, z_t) = \omega\big(\varphi(z_1, a_1, \ldots, z_{t-1}, a_{t-1}, z_t)\big) \quad \text{and}$$
$$\varphi(z_1, a_1, \ldots, z_t, a_t, z_{t+1}) = g\big(\varphi(z_1, a_1, \ldots, z_{t-1}, a_{t-1}, z_t), a_t, z_{t+1}\big).$$

As σ^1 is φ-factored and φ is recursive, $V_\beta^2(\sigma^1 \mid h) = V_\beta^2(\sigma^1 \mid h')$ whenever $\varphi(h) = \varphi(h')$. Let τ be a pure strategy of Player 2 such that $\tau(h) = \tau(h')$ whenever $\varphi(h) = \varphi(h')$ and $\tau(h)$ maximizes the left-hand side of (3). The existence of such a strategy τ follows from the facts that φ is recursive and that there are only finitely

Journal: 182 Article No.: 0733 TYPESET DISK LE CP Disp.:2020/10/1 Pages: 20 Layout: Small-Ex

many possible actions for Player 2 in each state. Hence, τ is a φ-factored strategy. For every $h = (z_1, a_1, \ldots, z_t)$ and every strategy σ^2 we have

$$\begin{aligned} E_{(\sigma^1,\tau|h)}\Big(u_2(z_t,a_t)+\beta V_\beta^2\left(\sigma^1 \mid h,a_t,z_{t+1}\right)\Big) &\geq E_{(\sigma^1,\sigma^2|h)}\Big(u_2(z_t,a_t)+\beta V_\beta^2\left(\sigma^1 \mid h,a_t,z_{t+1}\right)\Big) \\ &\geq E_{(\sigma^1,\sigma^2|h)}\big(u_2(z_t,a_t)+\beta V_\beta^2(\sigma^1,\sigma^2 \mid h,a_t,z_{t+1})\big) \\ &= V_\beta^2\left(\sigma^1,\sigma^2 \mid h\right). \end{aligned}$$

This implies that

$$E_{(\sigma^1,\tau|h)}\Big(u_2(z_t,a_t)+\beta V_\beta^2\left(\sigma^1 \mid h,a_t,z_{t+1}\right)\Big) \geq V_\beta^2\left(\sigma^1 \mid h\right). \tag{4}$$

Since the opposite inequality in (4) clearly holds we conclude that τ satisfies (3). Therefore τ is a best reply to σ^1 in the β-discounted games $(\Gamma \mid h)$. This completes the proof of Theorem 4.1.

Proof of Theorem 4.2

In the forthcoming lemmas we fix a stochastic game Γ with finitely many states and actions and assume that φ is a recursive factor with a finite range X. Let H denote the set of all finite histories $h = (z_1, a_1, \ldots, z_t)$ of the stochastic game.

Let $\mathbb{R}^X$ be the finite-dimensional space of all real-valued functions defined on X with the maximum norm, and let $\ell_\infty(H)$ be the Banach space of all bounded real-valued functions defined on H with the supremum norm. Let σ^1 and σ^2 be strategies of Player 1 and Player 2, respectively. Define the map $T_\beta\colon \ell_\infty(H) \to \ell_\infty(H)$ by

$$(T_\beta V)(h) = E_{(\sigma^1,\sigma^2|h)}\big(u_2(z(h),a_t)+\beta V(h,a_t,z_{t+1})\big),$$

where $h = (z_1, a_1, \ldots, z_t)$ is a finite history. Then, T_β is a strict contraction ($\|T_\beta V - T_\beta V'\| \leq \beta\|V - V'\|$) and hence has a unique fixed point $V_\beta \in \ell_\infty(H)$. As

$$V_\beta^2(\sigma^1,\ \sigma^2 \mid h) = E_{(\sigma^1,\sigma^2|h)}\Big(u_2(z(h),a_t)+\beta V_\beta^2\left(\sigma^1,\ \sigma^2 \mid h,a_t,z_{t+1}\right)\Big),$$

we get $V_\beta(h) = V_\beta^2(\sigma^1,\ \sigma^2 \mid h)$.

Assume that σ^1 and σ^2 are φ-factored strategies of Player 1 and Player 2, respectively. As φ is recursive, the map $T\colon \mathbb{R}^X \to \mathbb{R}^X$ that is defined by

$$(TU)(x) = E_{(\sigma^1,\sigma^2|h)}\Big(u_2\big(z(h),a_t\big)+\beta U(\varphi(h,a_t,z_{t+1}))\Big),$$

where $h = (z_1, a_1, \ldots, z_t)$ is any finite history with $\varphi(h) = x$, is well defined. In addition, T is a strict contraction ($\|TU - TU'\| \leq \beta\|U - U'\|$), and hence has a unique fixed point $U_\beta \in \mathbb{R}^X$.

Journal: 182 Article No.: 0733 TYPESET DISK LE CP Disp.:2020/10/1 Pages: 20 Layout: Small-Ex

Note that $V \in \ell_\infty(H)$ that is defined by $V(h) = U_\beta(\varphi(h))$ is a fixed point of T_β. Hence, $V^2_\beta(\sigma^1, \sigma^2 \mid h)$ equals $U_\beta(\varphi(h))$ and is a function of β, σ^1, σ^2, and $\varphi(h)$.

Lemma 4.4 *For every finite history h and φ-factored strategies σ^1 of Player 1 and σ^2 of Player 2, $V^2_\beta(\sigma^1, \sigma^2 \mid h)$ is a rational function of $\beta \in (0, 1)$.*

Proof Recall that a factor φ in a stochastic game determines the current state. The state determined by $x \in X$ is denoted by $z(x)$. Therefore, if $h = (z_1, a_1, \ldots, z_t)$ and $\varphi(h) = x$ we have $z(x) = z_t$. Fix a pair of φ-factored strategies $\sigma^1 = \omega^1 \circ \varphi$ and $\sigma^2 = \omega^2 \circ \varphi$. Define the vector $u \in \mathbb{R}^X$ and the $X \times X$ transition matrix Q by

$$u(x) = E_{(\sigma^1, \sigma^2 \mid h)} u_2(z_t, a_t)$$

and

$$Q(x, x') = P_{(\sigma^1, \sigma^2 \mid h)}\big(\varphi(h, a_t, z_{t+1}) = x'\big),$$

where $h = (z_1, \ldots, z_t)$ with $\varphi(h) = x$. As the factor φ is recursive and σ^1 and σ^2 are φ-factored, u and Q are well defined. Note that the vector $U \in \mathbb{R}^X$ that is given by

$$U(x) = \sum_{x' \in X} \Big(u(x') \cdot \sum_{j=0}^{\infty} \beta^j Q^j(x, x') \Big), \qquad x \in X,$$

is a fixed point of T. As Q is a transition matrix, $\sum_{j=0}^{\infty} \beta^j Q^j = (I - \beta Q)^{-1}$ and

$$U(x) = \sum_{x' \in X} u(x') \cdot (I - \beta Q)^{-1}(x, x').$$

Hence,

$$V_\beta\left(\sigma^1, \sigma^2 \mid h\right) = U_\beta(\varphi(h)) = U(\varphi(h)) = \sum_{x' \in X} u(x')(I - \beta Q)^{-1}(\varphi(h), x').$$

Each one of the entries of $(I - \beta Q)^{-1}$ is a rational function of $\beta \in (0, 1)$. Hence, $V^2_\beta(\sigma^1, \sigma^2 \mid h)$ is a rational function of $\beta \in (0, 1)$.

Lemma 4.5 *For every $x \in X$ and a φ-factored strategy σ^1 of Player 1 there exists $\beta_1(x) \in (0, 1)$ such that if $\varphi(h) = x$ then $\beta \mapsto V^2_\beta(\sigma^1 \mid h)$ is a rational function of $\beta \in [\beta_1(x), 1)$.*

Proof Fix a finite history h and a φ-factored strategy σ^1 of Player 1. By Theorem 4.1, for every $\beta \in (0, 1)$, there is a pure φ-factored strategy τ of Player 2 such that $V^2_\beta(\sigma^1 \mid h) = V^2_\beta(\sigma^1, \tau \mid h)$. There are finitely many pure φ-factored strategies τ of Player 2 and for each such strategy τ, $V^2_\beta\left(\sigma^1, \tau \mid h\right)$ is a rational function of $\beta \in (0, 1)$ by Lemma 4.4. Hence, $V^2_\beta(\sigma^1 \mid h)$ is the maximal value of finitely many

Journal: 182 Article No.: 0733 TYPESET DISK LE CP Disp.:2020/10/1 Pages: 20 Layout: Small-Ex

rational functions of β. Two distinct rational functions coincide in at most finitely many points. Therefore, there is $\beta_*(h) \in (0,1)$ such that $V^2_\beta(\sigma^1 \mid h)$ is a rational function of $\beta \in (\beta_*(h), 1]$.

The range of φ is finite; therefore, there is $\beta_1(x) \in (0,1)$ such that, for every $h \in H$, $\beta \mapsto V^2_\beta\left(\sigma^1 \mid h\right)$ is a rational function of $\beta \in [\beta_1, 1)$.

Let $\sigma^1 = \omega^1 \circ \varphi$ be a φ-factored strategy, where $\varphi\colon H \to X$ is a recursive factor with a finite range. For every $\beta \in (0,1)$, $x \in X$, and $a^2 \in A_2(z(x))$, we define

$$M(\beta, x, a^2) = \sum_{a^1 \in A_1(z(x))} \Big(\omega^1(x)(a^1) \cdot \Big[u_2\big(z(x), (a^1, a^2)\big) + \beta \cdot \sum_{z \in S} \Big(p\big(z(x), (a^1, a^2)\big)(z) \cdot V^2_\beta\big(\sigma^1 \mid h, (a^1, a^2), z\big)\Big)\Big]\Big)$$

where $h \in H$ satisfies $\varphi(h) = x$. Since σ^1 is φ-factored and φ is recursive, the function M is well defined. Using Lemma 4.5 we see that the function $\beta \mapsto M(\beta, x, a^2)$ is a rational function of $\beta \in [\beta_1, 1)$, where β_1 is taken from Lemma 4.5 and x, a^2 are fixed. Hence, there is $\beta_0(x) \in (0,1)$ and $\omega^2(x) \in A_2(z(x))$ such that, for every $\beta \in [\beta_0(x), 1)$, we have

$$M\big(\beta, x, \omega^2(x)\big) = \max_{a^2 \in A(z(x))} M\left(\beta, x, a^2\right).$$

By definition we also get $M\big(\beta, x, \omega^2(x)\big) = V^2_\beta(\sigma^1 | h)$ for every $x \in X$, $h \in H$ with $\varphi(h) = x$, and $\beta \in [\beta_0(x), 1)$. By setting $\beta_0 = \max_{x \in X} \beta_0(x)$, for every $x \in X$ and $\beta \in [\beta_0, 1)$, we have

$$M\big(\beta, x, \omega^2(x)\big) = \max_{a^2 \in A(z(x))} M\left(\beta, x, a^2\right).$$

Therefore, for every discount factor $\beta \in [\beta_0, 1)$ and $h \in H$, we have $\omega^2 \circ \varphi$ is a best reply to σ^1 in the β-discounted games $(\Gamma \mid h)$.

Proof of Theorem 4.3

Let $\varphi\colon H \to X$ be a recursive factor with a finite range and σ^1 be a φ-factored strategy of Player 1. Let σ^2 be a φ-factored strategy of Player 2 and $\beta_0 \in (0,1)$ be a discount factor such that for every discount factor $\beta \in [\beta_0, 1)$ and history h, σ^2 is a best reply to σ^1 in the β-discounted game $(\Gamma \mid h)$; that is, $V^2_\beta(\sigma^1, \sigma^2 \mid h) \geq V^2_\beta(\sigma^1, \rho \mid h)$ for every strategy ρ of Player 2. The existence of such σ^2 and β_0 follows from Theorem 4.2.

Set $x_t = \varphi(z_1, a_1, \ldots, z_t)$ and $v_\lambda(x) := \lambda V^2_{1-\lambda}(\sigma^1, \sigma^2 \mid h)$ whenever $\varphi(h) = x$. As σ^1 and σ^2 are φ-factored and φ is a recursive factor, $v_\lambda(x)$ is well defined. The function $(0,1] \ni \lambda \mapsto v_\lambda(x)$ is a bounded rational function. Hence, it is a Lipschitz function of λ. Let K be a sufficiently large positive constant such that $|v_\lambda(x) - v_{\lambda'}(x)| \leq K|\lambda - \lambda'|$ for every $x \in X$. Note that $v_\infty(x) := \lim_{\lambda \to 0+} v_\lambda(x)$ exists.

Journal: 182 Article No.: 0733 TYPESET DISK LE CP Disp.:2020/10/1 Pages: 20 Layout: Small-Ex

For $1 > \beta \geq \beta_0$, $V_\beta^2(\sigma^1, \sigma^2 \mid h) = V_\beta^2(\sigma^1 \mid h)$, and hence, for $h = (z)$ and for every $0 < \lambda \leq 1 - \beta_0$, we have

$$E_{(\sigma^1,\sigma^2|z)}\big(\lambda u_2(z_t, a_t) + (1-\lambda)v_\lambda(x_{t+1}) \mid \mathcal{H}_t\big) \geq v_\lambda(x_t), \tag{5}$$

where $\mathcal{H}_t$ is the algebra of events defined by histories up to stage t. For any strategy ρ of Player 2 and every $0 < \lambda \leq 1 - \beta_0$, we have

$$E_{(\sigma^1,\rho|z)}\big(\lambda u_2(z_t, a_t) + (1-\lambda)v_\lambda(x_{t+1}) \mid \mathcal{H}_t\big) \leq v_\lambda(x_t). \tag{6}$$

The continuation of the proof uses the proof of Mertens and Neyman (1981). Let A be the largest absolute value of a payoff appearing in the game matrices. Note that inequality (5) is identical to inequality (2.1) of Mertens and Neyman (1981), and inequality (6) is a dual version of inequality (2.1) of Mertens and Neyman (1981). The integrable function $\psi : (0, 1] \to \mathbb{R}$ that satisfies inequality (2.2) of Mertens and Neyman (1981) is simply the constant function K.

As in the proof of Mertens and Neyman (1981), for every $\varepsilon > 0$ and $1 > \beta^* \geq \beta_0$, there are

- a sequence $(\beta_t)_{t=1}^\infty$, where β_t is a function of the history $(z_1, a_1, \ldots, z_t)$ with $\beta_t \geq \beta^*$ and
- a constant s_0,

such that the following conditions are satisfied.

(a) We denote $\lambda_t := 1 - \beta_t$. Then we have $\lim_{t\to\infty} \lambda_t = 0$ $P_{(\sigma^1,\sigma^2|z)}$-a.e.
(b) Let I be the indicator function. Then $E_{(\sigma^1,\sigma^2|z)}\big(\sum_{t=1}^\infty I(\beta_t = \beta^*)\big) \leq \frac{2A}{\varepsilon(1-\beta^*)}$.
(c) We have

$$E_{(\sigma^1,\sigma^2|z)}\big(v_{\lambda_t}(x_{t+1})\big) \geq v_\infty(\varphi(z)) - 3\varepsilon.$$

(d) There exists a function Y_∞ such that $v_{\lambda_t}(x_{t+1}) \to Y_\infty$ $P_{(\sigma^1,\sigma^2|z)}$-a.e., and $E_{(\sigma^1,\sigma^2|z)}(Y_\infty \mid z) \geq v_\infty(\varphi(z)) - 2\varepsilon$.
(e) For every n we have

$$\sum_{t<n} u_2(z_t, a_t) \geq \sum_{t<n} v_{\lambda_t}(x_{t+1}) - s_0 - 2A \sum_{t<n} I(\beta_t = \beta^*) - 4n\varepsilon.$$

Therefore, we have

$$\begin{aligned} E_{(\sigma^1,\sigma^2|z)}\left(\liminf_{n\to\infty} \frac{1}{n}\sum_{t=1}^n u_2(z_t, a_t)\right) &\geq E_{(\sigma^1,\sigma^2|z)}\left(\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^n v_{\lambda_t}(x_{t+1})\right) - 4\varepsilon \\ &= E_{(\sigma^1,\sigma^2|z)}(Y_\infty \mid z) - 4\varepsilon \geq v_\infty(\varphi(z)) - 6\varepsilon. \end{aligned}$$

Journal: 182 Article No.: 0733 TYPESET DISK LE CP Disp.:2020/10/1 Pages: 20 Layout: Small-Ex

Further, for n sufficiently large such that $\frac{s_0}{n} + \frac{2A}{n}\frac{2A}{\varepsilon(1-\beta^*)} < \varepsilon$, we have

$$E_{(\sigma^1,\sigma^2|z)}\left(\frac{1}{n}\sum_{t=1}^{n} u_2(z_t,a_t)\right) \geq E_{(\sigma^1,\sigma^2|z)}\left(\frac{1}{n}\sum_{t=1}^{n} v_{\lambda_t}(z_{t+1})\right) - 4\varepsilon$$
$$\geq v_\infty(\varphi(z)) - 5\varepsilon.$$

Similarly, if ρ is a strategy of player 2, then

$$E_{(\sigma^1,\rho|z)}\left(\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n} u_2(z_t,a_t)\right) \leq E_{(\sigma^1,\sigma^2|z)}\left(\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n} v_{\lambda_t}(x_{t+1})\right) + 4\varepsilon$$
$$= E_{(\sigma^1,\sigma^2|z)}(Y_\infty \mid z) + 4\varepsilon \leq v_\infty(\varphi(z)) + 6\varepsilon,$$

and for n sufficiently large such that $\frac{s_0}{n} + \frac{2A}{n}\frac{2A}{\varepsilon(1-\beta^*)} < \varepsilon$,

$$E_{(\sigma^1,\rho|z)}\left(\frac{1}{n}\sum_{t=1}^{n} u_2(z_t,a_t)\right) \leq E_{(\sigma^1,\sigma^2|z)}\left(\frac{1}{n}\sum_{t=1}^{n} v_{\lambda_t}(x_{t+1})\right) + 4\varepsilon$$
$$\leq v_\infty(\varphi(z)) + 5\varepsilon.$$

The last four displayed inequalities prove (1) and (2) for $h = z$. The proof of inequalities (1) and (2) for a general finite history $h = (z_1, a_1, \ldots, z_{t'})$ is the same: just replace, in the above inequalities, $P_{(\sigma^1,\sigma^2|z)}$ by $P_{(\sigma^1,\sigma^2|h)}$, $P_{(\sigma^1,\rho|z)}$ by $P_{(\sigma^1,\rho|h)}$, n by $t' + n$, and $t \in \mathbb{N}$ by $t \geq t'$.

Remark 4.6 An alternative proof of Theorem 4.3 that does not rely on the Mertens–Neyman proof (Mertens and Neyman (1981)) follows the proof of parts 3) and 4) of [Neyman (2003) Proposition 3].

5 Corollaries of the main result

The next corollary follows from the facts that a supergame and a supergame with time-dependent stage games are special cases of a stochastic game and that a behavioral k-SBR, a behavioral time-dependent $k(t)$-BR strategy, a behavioral k-state automaton strategy, and a behavioral k-state time-dependent automaton strategy are factored strategies with the corresponding natural recursive factors. Note that the results for supergames and strategies with a finite history can be found in Press and Dyson (2012).

Corollary 5.1 *Fix a discount factor $\beta \in (0,1)$. Let an infinite-stage game be either a supergame, or a time-dependent supergame, or a stochastic game with countably many states. If the stage games have finitely many actions, then the β-discounted infinite-stage game obeys the following best-reply properties.*

(a) *Let k be a nonnegative integer. Then, for every behavioral k-SBR strategy σ^1, there is a pure k-SBR strategy σ^2 that is a best reply of Player 2.*

Journal: 182 Article No.: 0733 TYPESET DISK LE CP Disp.:2020/10/1 Pages: 20 Layout: Small-Ex

(b) *Let $k\colon \mathbb{N} \to \mathbb{N}$ be a function with $k(t+1) \leq k(t)+1$ for every $t \in \mathbb{N}$. Then, for every behavioral time-dependent k-BR strategy σ^1, there is a pure k-BR strategy σ^2 that is a best reply.*

(c) *Let k be a positive integer. Then, for every strategy σ^1 that is defined by a k-state behavioral automaton, there is a best reply σ^2 of Player 2 that is defined by a deterministic k-state automaton.*

(d) *Let k be a positive integer. Then, for every strategy σ^1 that is defined by a k-state time-dependent automaton, there is a best reply σ^2 that is defined by a deterministic k-state time-dependent automaton.*

The next two theorems are well-known results in the theory of Markov decision processes. The first is analogous to Theorem 4.1 and the second is analogous to Theorem 4.2. As an MDP is a one-person stochastic game, these results follow from our main result. An MDP is defined by a tuple $\langle S, A, u, p, \mu\rangle$, where S is the state space, A is the set of actions of the decision maker (DM), u is the payoff to the DM, p is the transition probability, and μ is the probability of the initial state. If σ is a strategy of DM, then $v_\beta(\sigma)$ denotes the β-discounted payoff of DM as a function of the strategy σ.

Theorem 5.2 (Derman 1965) *Let $\langle S, A, u, p, \mu\rangle$ be an MDP with countably many states, finitely many actions in each state, and a bounded reward function. Then for each $\beta \in (0,1)$ there is a stationary pure strategy σ such that, for every strategy ρ, we have $v_\beta(\sigma) \geq v_\beta(\rho)$.*

Theorem 5.3 (Blackwell 1962) *Let $\langle S, A, u, p, \mu\rangle$ be an MDP with finitely many states and actions. Then there is a stationary pure strategy σ and $\beta_0 \in (0,1)$ such that, for every strategy ρ and for every $\beta \in [\beta_0, 1)$, we have $v_\beta(\sigma) \geq v_\beta(\rho)$.*

Remark 5.4 Theorem 4.1 and Theorem 4.2 are analogous to Theorem 5.2 and Theorem 5.3 respectively. A natural approach is to try to deduce Theorems 4.1 and 4.2 from these theorems by assigning to a φ-factored strategy σ^1 in a two-person stochastic game with countably many states and finitely many actions the following auxiliary MDP. The state space of the auxiliary MDP is X, where X is the image of the recursive factor φ. The action sets of DM are those of Player 2 in the stochastic game. The stage payoffs are defined by

$$u(x, a^2) = \sum_{a^1 \in A_1} \omega^1(x)\left(a^1\right) \cdot u_2\left(a^1, a^2\right), \quad x \in X \text{ and } a^2 \in A_2,$$

where $\omega^1 \circ \varphi$ is the φ-factored strategy of Player 1, and the transitions are defined by

$$p\left(x, a^2\right)(x') = \sum_{a^1 \in A_1} \omega^1(x)\left(a^1\right) \cdot p\left(x, a^1, a^2\right)(x'), \quad x, x' \in X \text{ and } a^2 \in A_2.$$

By Theorem 5.2 the decision maker has a pure stationary optimal strategy σ in the auxiliary β-discounted MDP, and a stationary strategy in the auxiliary MDP maps naturally to a φ-factored strategy in the stochastic game.

Hence, it is tempting to argue that this proves that this φ-factored strategy in the stochastic game is a best reply of Player 2 to the φ-factored strategy σ^1 of Player 1.

However, in the stochastic game, a strategy of Player 2 may depend also on the past actions of Player 1. In the mapping of the stochastic game to the auxiliary MDP the history of past actions of Player 1 disappears. Hence, the φ-factored strategy σ is a best reply among all strategies that ignore past actions of Player 1, but, at least formally, one cannot argue that it is a best reply among all strategies of Player 2.

6 Remarks

6.1 The role of recursiveness

Recursiveness of φ is not a necessary condition in the main results, even in the model of the supergame. For example, consider a supergame, where each player has at least two actions, and let φ be a nonrecursive factor that is independent of the actions of Player 2. Then, for every action function $\omega^1 \colon X \to \Delta(A_1)$ of Player 1, if $\omega^2 \colon X \to \Delta(A_2)$ is the best reply to ω^1, i.e., $u_2\big(\omega^1(x), \omega^2(x)\big) \geq u_2\big(\omega^1(x), a^2\big)$ for every $x \in X, a^2 \in A_2$, then the φ-factored strategy $\omega^2 \circ \varphi$ of Player 2 is a best reply to the φ-factored strategy $\omega^1 \circ \varphi$ of Player 1 in any β-discounted game $(G \mid h)$.

However, if the factor φ is nonrecursive and obeys an additional condition that is spelled out below, then the conclusion of Theorem 4.1 does not hold even in the model of the discounted supergame. We illustrate this remark in the model of the supergame where each player has at least two actions. Without loss of generality we can assume that $A_i = \{0, 1\}$ and that the factor $\varphi \colon H \to X$ obeys $\varphi(h) = \varphi(h')$, $\varphi(h, (1, 1)) \neq \varphi(h', (1, 1))$ for some histories h, h'. The added condition is

$$\big\{\varphi(h), \varphi\big(h, (1, 1)\big), \varphi\big(h', (1, 0)\big)\big\} \cap \big\{\varphi\big(h', (1, 1)\big), \varphi\big(h, (1, 0)\big)\big\} = \emptyset. \tag{7}$$

Then we can find

- a payoff function $u_2 \colon A_1 \times A_2 \to [0, 1]$ and
- a pure action function $\omega^1 \colon X \to A_1$,

such that any strategy τ^2 of Player 2 that is a best reply to $\omega \circ \varphi$ in both of the β-discounted games, $(G \mid h)$ and $(G \mid h')$, obeys $\tau^2(h) \neq \tau^2(h')$; thus τ^2 is not φ-factored.

Using (7) we can define a φ-factored strategy $\omega^1 \circ \varphi$ of Player 1 by setting

$$\omega^1\big(\varphi(h)\big) = \omega^1\big(\varphi(h')\big) = \omega^1\big(\varphi(h, (1, 1))\big) = \omega^1\big(\varphi(h', (1, 0))\big) = 1 \quad \text{and}$$
$$\omega^1\big(\varphi(h', (1, 1))\big) = \omega^1\big(\varphi(h, (1, 0))\big) = 0.$$

Let the payoff function u_2 of Player 2 be given by

$$u_2(1, \cdot) = 1 \text{ and } u_2(0, \cdot) = 0.$$

Journal: 182 Article No.: 0733 TYPESET DISK LE CP Disp.:2020/10/1 Pages: 20 Layout: Small-Ex

We will deal with the supergame whose stage game G is $\langle A_1, A_2, u_1, u_2\rangle$. Fix a sufficiently small discount factor $\beta \in (0, 1)$ such that $\beta > \beta^2/(1-\beta)$. Consider the supergame $(G \mid h)$, i.e., the supergame that follows history h. If Player 2 plays action 1 in the first stage of $(G \mid h)$, then she guarantees that her (unnormalized) β-discounted payoff is least $1+\beta$. If Player 2 plays action 0 in the first stage of the supergame $(G \mid h)$, then her (unnormalized) β-discounted payoff is at most $1+\beta^2/(1-\beta)$. Therefore, any strategy of Player 2 in which action 1 is played in the first stage of $(G \mid h)$ dominates any strategy of Player 2 in which action 0 is played in the first stage of $(G \mid h)$. Therefore, if σ^2 is a best reply of Player 2 to $\omega^1 \circ \varphi$ in the β-discounted game $(G \mid h)$, then $\sigma^2(h) = 1$.

Similarly, consider the supergame $(G \mid h')$. If Player 2 plays action 0 in the first stage of $(G \mid h')$, then she guarantees that her (unnormalized) β-discounted payoff is at least $1+\beta$. If Player 2 plays action 1 in the first stage of $(G \mid h')$, then her (unnormalized) β-discounted payoff is at most $1+\beta^2/(1-\beta)$. Therefore any strategy of Player 2 in which action 0 is played in the first stage of $(G \mid h')$ dominates any strategy of Player 2 in which action 1 is played in the first stage of $(G \mid h')$. Therefore, if σ^2 is a best reply of Player 2 to $\omega^1 \circ \varphi$ in the β-discounted game $(G \mid h')$, then $\sigma^2(h') = 0$.

Hence, if σ^2 is a best reply of Player 2 to $\omega^1 \circ \varphi$ in each one of the β-discounted games $(G \mid h)$ and $(G \mid h')$, then $\sigma^2(h) \neq \sigma^2(h')$ and, therefore, as $\varphi(h) = \varphi(h')$, σ^2 is not φ-factored.

6.2 Compact action spaces

A natural extension of our model is to consider players with compact action sets A_i. In this extension, there arises a new problem not found in games with finite action profiles, namely, the existence of a best reply to a given strategy σ. Consider, for example, the following two-player supergame, where the set of actions of each player is the interval $[0, 1]$ and the stage payoff of Player 2 is (at any time) given by $u_2(a^1, a^2) = a^1 + a^2$. Now, suppose that Player 1 plays the 1-SBR strategy given by

$$\sigma^1(a_1, \ldots, a_{t-1}) = \begin{cases} 1, & \text{if } a_{t-1}^2 < 1 \text{ and } t > 1, \\ 0, & \text{otherwise,} \end{cases} \qquad e_1 = (0, 0).$$

The strategy σ^1 is a factor-based strategy with a recursive factor. Indeed, we set $X = \{B, C\}$, $\omega(B) = 1$ and $\omega(C) = 0$, and $\varphi(a_1, \ldots, a_{t-1}) = B$ if $a_{t-1}^2 < 1$ and $t > 1$, $\varphi(a_1, \ldots, a_{t-1}) = C$. Then we have $\sigma^1 = \omega \circ \varphi$. However, in the β-discounted game there is no φ-factored best reply, and any φ-factored reply is dominated by (another) φ-factored reply.

Of course, there does not exist any general best reply to σ^1. The difficulty stems from the fact that the factor φ is not continuous. However, in analogy to, and using the tools of, say, Maitra (1968), one can generalize Theorem 4.1 to the setup with a continuous factor.

Journal: 182 Article No.: 0733 TYPESET DISK LE CP Disp.:2020/10/1 Pages: 20 Layout: Small-Ex

6.3 Public vs. private strategies

A natural recursive factor φ arises in the theory of repeated games with imperfect monitoring. This natural factor φ maps any finite history of the game to the finite history x of the public signals, and we can identify the φ-factored strategies with the so-called *public strategies* [see, e.g., (Radner et al. 1986)]. By contrast, *a private strategy* [see, e.g.,(Kandori and Obara 2006)] is a strategy where the current action depends on the history of public signals (i.e., on elements of X), as well as on the history of private signals (e.g., past own actions).

Our question at the outset of this paper can then be reformulated to "If my opponent is using a public strategy, when can I benefit by using my (additional) private signals?"; in other words, "Can a private strategy fare better than any public strategy against a public strategy?" The answer is that in the discounted repeated game with imperfect monitoring (finitely many stage actions and countably many public signals) one does not profit from the additional private signals since the factor φ is obviously recursive.

References

Abreu D, Rubinstein A (1988) The structure of Nash equilibrium in repeated games with finite automata. Econometrica 56(6):1259–1281

Aumann RJ (1981) Survey of repeated games. In: Essays in game theory and mathematical economics in Honour of Oskar Morgenstern, pp 11–42. Wissenschaftsverlag, Bibliographisches Institut

Aumann RJ, Sorin S (1989) Cooperation and bounded recall. Games Econ Behavior 1(1):5–39

Ben-Porath E (1993) Repeated games with finite automata. J Econ Theory 59:17–32

Blackwell D (1962) Discrete dynamic programing. Ann Math Stat 33:719–726

Derman C (1965) Markovian sequential control processes: denumerable state spaces. J Math Anal Appl 10:295–302

Kalai E (1990) Bounded rationality and strategic complexity in repeated games. In: Ichiishi T, Neyman A, Tauman Y (eds) Game theory and applications. Academic Press, San Diego, pp 131–157

Kalai E, Stanford W (1988) Finite rationality and interpersonal complexity in repeated games. Econometrica 56(2):397–410

Kandori M, Obara I (2006) Efficiency in repeated games revisited: the role of private strategies. Econometrica 74(2):499–519

Lehrer E (1988) Repeated games with stationary bounded recall strategies. J Econ Theory 46(1):130–144

Maitra A (1968) Discounted dynamic programming on compact metric spaces. Sankhyā Indian J Stat Ser A 30(2):211–216

Maskin E, Tirole J (2001) Markov perfect equilibrium. I. Observable actions. J Econ Theory 100(2):191–219

Mertens J-F, Neyman A (1981) Stochastic games. Internat J Game Theory 10(2):53–66

Neyman A (1997) Cooperation, repetition, and automata. In: Hart S, Mas-Colell A (eds) Cooperation: game theoretic approaches, volume 155 of NATO ASI Series F, pp 233–255

Neyman A (1985) Bounded complexity justifies cooperation in the finitely repeated prisoners' dilemma. Econ Lett 19:227–229

Neyman A (2003) From Markov chains to stochastic games. In: Neyman A, Sorin S (eds) Stochastic games and applications. Springer, Dordrecht, pp 9–25

Neyman A, Okada D (2009) Growth of strategy sets, entropy, and nonstationary bounded recall. Games Econ Behavior 66(1):404–425

Press WH, Dyson FJ (2012) Iterated prisoner's dilemma contains strategies that dominate any evolutionary opponent. PNAS 109(26):10409–10413

Radner R, Myerson R, Maskin E (1986) An example of a repeated partnership game with discounting and with uniformly inefficient equilibria. Rev Econ Stud 53(172):59–69

Rubinstein A (1986) Finite automata play the repeated prisoner's dilemma. J Econ Theory 39(1):83–96

Journal: 182 Article No.: 0733 TYPESET DISK LE CP Disp.:2020/10/1 Pages: 20 Layout: Small-Ex

Affiliations

René Levínský[1] · Abraham Neyman[2] · Miroslav Zelený[3]

✉ René Levínský
rene.levinsky@cerge-ei.cz

Abraham Neyman
aneyman@math.huji.ac.il

Miroslav Zelený
zeleny@karlin.mff.cuni.cz

[1] Economics Institute of the Czech Academy of Sciences, Politických vězňů 7, 111 21 Praha 1, Czech Republic

[2] Institute of Mathematics and Center for the Study of Rationality, The Hebrew University of Jerusalem, Givat Ram, 91904 Jerusalem, Israel

[3] Charles University, Faculty of Mathematics and Physics, Sokolovská 83, 186 75 Praha 8, Czech Republic