Strategy Improvement for Concurrent Reachability Games

Krishnendu Chatterjee  Luca de Alfaro  Thomas A. Henzinger
UC Berkeley  UC Santa Cruz  EPFL and UC Berkeley

Abstract

A concurrent reachability game is a two-player game played on a graph: at each state, the players simultaneously and independently select moves; the two moves determine jointly a probability distribution over the successor states. The objective for player 1 consists in reaching a set of target states; the objective for player 2 is to prevent this, so that the game is zero-sum.

Our contributions are two-fold. First, we present a simple proof of the fact that in concurrent reachability games, for all $\varepsilon > 0$, memoryless $\varepsilon$-optimal strategies exist. A memoryless strategy is independent of the history of plays, and an $\varepsilon$-optimal strategy achieves the objective with probability within $\varepsilon$ of the value of the game. In contrast to previous proofs of this fact, which rely on the limit behavior of discounted games using advanced Puisieux series analysis, our proof is elementary and combinatorial. Second, we present a strategy-improvement (a.k.a. policy-iteration) algorithm for concurrent games with reachability objectives.

1. Introduction

We consider concurrent reachability games played by two players on finite state spaces. The configuration of such a game is called a state. At each round, the two players choose their moves concurrently and independently; the two moves and the current state determine a successor state, or in general, a probability distribution over the successor states. A play of the game consists in the infinite sequence of states visited while playing the game. The objective of player 1 consists in forcing the game to a specified set of target states; the objective of player 2 consists in preventing the game from reaching a target state. Consequently, we assign value 1 to all plays that reach the target set, and value 0 to all other plays. The players can adopt strategies that are both randomized and history-dependent. Player 1 can guarantee a value $v$ for the game from a state $s$ if player 1 has a strategy that ensures that the expected value of a play from $s$ is at least $v$, regardless of the strategy chosen by player 2. The value at state $s$ of the reachability game with target $T$ is the supremum of the set of values that player 1 can guarantee from $s$. An optimal strategy for player 1 is a strategy that guarantees the value of the game from each state $s$. For $\varepsilon > 0$, an $\varepsilon$-optimal strategy for player 1 is a strategy that guarantees that the objective is satisfied with a probability within $\varepsilon$ of the value of the game, for each state $s$.

Concurrent reachability games belong to the family of repeated games [17, 13], and they have been studied more specifically in [9, 8, 10]. In this paper our contributions are two-fold. First, we present a simple and combinatorial proof of the existence of memoryless $\varepsilon$-optimal strategies for concurrent games with reachability objectives, for all $\varepsilon > 0$. Second, we present a strategy-improvement (a.k.a. policy-iteration) algorithm for concurrent reachability games. Unlike in the special case of turn-based games, where at every state at most one player can choose between multiple moves, the algorithm need not terminate in finitely many iterations. Strategy improvement algorithms were previously known for turn-based games with reachability objectives [5], and turn-based games with more complex objectives [18, 2].

It has long been known that optimal strategies need not exist for concurrent reachability games [13], so that one must settle for $\varepsilon$-optimality. It was also known that, for $\varepsilon > 0$, there exist $\varepsilon$-optimal strategies that are memoryless, i.e., strategies that always choose a probability distribution over moves that depends only on the current state, and not on the past history of the play [14]. Unfortunately, the only previous proof of this fact is rather complex. The proof considered discounted versions of reachability games, where a play that reaches the target in $k$ steps is assigned a value of $\alpha^k$, for some discount factor $0 < \alpha \leq 1$, rather than value 1. It is possible to show that, for $0 < \alpha < 1$, memoryless optimal strategies always exist. The result for the undiscounted ($\alpha = 1$) case followed from an analysis of the limit behavior of such optimal strategies for $\alpha \to 1$. The limit behavior is studied with the help of results on the field of real Puisieux series [14]. This proof idea works not only for reachability games, but also for total-reward games with nonnegative rewards (see [14] again). A more specialized recent result [12] established the existence of memoryless $\varepsilon$-optimal strategies for certain infinite-state (recursive) concurrent games, but again the proof relies on deep results from analysis and
linear algebra (matrix theory). We show that the existence of memoryless \( \varepsilon \)-optimal strategies for concurrent reachability games can be established by more elementary means. Our proof relies only on combinatorial techniques and on simple properties of Markov decision processes \([1, 7]\). As our proof is easily accessible, we believe that the proof techniques we use will find future applications in game theory.

Our proof of the existence of memoryless \( \varepsilon \)-optimal strategies, for all \( \varepsilon > 0 \), is built upon a value-iteration scheme that converges to the value of the game \([10]\). The value-iteration scheme computes a sequence \( u_0, u_1, u_2, \ldots \) of valuations, where for \( i = 0, 1, 2, \ldots \) each valuation \( u_i \) associates with each state \( s \) of the game a lower bound \( u_i(s) \) on the value of the game, such that \( \lim_{i \to \infty} u_i(s) \) converges to the value of the game at \( s \). From each valuation \( u_i \), we can easily extract a memoryless, randomized player-1 strategy, by considering the (randomized) choice of moves for player 1 that achieves the maximal one-step expectation of \( u_i \). In general, a strategy \( \pi_i \), obtained in this fashion is not guaranteed to achieve the value \( u_i \). We show that \( \pi_i \) is guaranteed to achieve the value \( u_i \) if it is proper, that is, if regardless of the strategy adopted by player 2, the game reaches with probability 1 states that are either in the target, or that have no path leading to the target. Next, we show how to extract from the sequence of valuations \( u_0, u_1, u_2, \ldots \) a sequence of memoryless randomized player-1 strategies \( \pi_0, \pi_1, \pi_2, \ldots \) that are guaranteed to be proper, and thus achieve the values \( u_0, u_1, u_2, \ldots \). This proves the existence of memoryless \( \varepsilon \)-optimal strategies for all \( \varepsilon > 0 \).

We then apply the techniques developed for the above proof to develop a strategy-improvement algorithm for concurrent reachability games. Strategy-improvement algorithms, also known as policy iteration algorithms in the context of Markov decision processes \([11, 1]\), compute a sequence of memoryless strategies \( \pi_0, \pi_1, \pi_2, \ldots \) such that, for all \( k \geq 0 \), (i) the strategy \( \pi_{k+1} \) is at all states no worse than \( \pi_k \); (ii) if \( \pi_{k+1} = \pi_k \), then \( \pi_k \) is optimal; and (iii) for every \( \varepsilon > 0 \), we can find a \( k \) sufficiently large so that \( \pi_k \) is \( \varepsilon \)-optimal. Computing a sequence of strategies \( \pi_0, \pi_1, \pi_2, \ldots \) on the basis the value-iteration scheme from above does not yield a strategy-improvement algorithm, as condition (ii) may be violated: there is no guarantee that a step in the value iteration leads to an improvement in the strategy. We will show that the key to obtain a strategy-improvement algorithm consists in recomputing, at each iteration, the values of the player-1 strategy to be improved, and in adopting a particular strategy-update rule, which ensures that all the strategies produced are proper. Unlike previous proofs of strategy-improvement algorithms for concurrent games \([5, 14]\), which relied on the analysis of discounted versions of the games, our analysis is again purely combinatorial. Differently from turn-based games \([5]\), for concurrent games we cannot guarantee the termination of the strategy-improvement algorithm. In fact, there are games where optimal strategies do not exist, and we can guarantee the existence of only \( \varepsilon \)-optimal strategies, for all \( \varepsilon > 0 \) \([13, 9]\).

2. Definitions

Notation. For a countable set \( A \), a probability distribution on \( A \) is a function \( \delta : A \to [0, 1] \) such that \( \sum_{a \in A} \delta(a) = 1 \). We denote the set of probability distributions on \( A \) by \( \mathcal{D}(A) \). Given a distribution \( \delta \in \mathcal{D}(A) \), we denote by \( \text{Supp}(\delta) = \{ x \in A | \delta(x) > 0 \} \) the support set of \( \delta \).

Definition 1 (Concurrent games) A (two-player) concurrent game structure \( G = \langle S, M, \Gamma_1, \Gamma_2, \delta \rangle \) consists of the following components:

- A finite state space \( S \) and a finite set \( M \) of moves.
- Two move assignments \( \Gamma_1, \Gamma_2 : S \to 2^M \setminus \emptyset \). For \( i \in \{1, 2\} \), assignment \( \Gamma_i \) associates with each state \( s \in S \) a nonempty set \( \Gamma_i(s) \subseteq M \) of moves available to player \( i \) at state \( s \).
- A probabilistic transition function \( \delta : S \times M \times M \to \mathcal{D}(S) \) that gives the probability \( \delta(s, a_1, a_2)(t) \) of a transition from \( s \) to \( t \) when player 1 chooses at state \( s \) move \( a_1 \) and player 2 chooses move \( a_2 \), for all \( s, t \in S \) and \( a_1 \in \Gamma_1(s), a_2 \in \Gamma_2(s) \).

We denote by \( |\delta| = \sum_{s \in S} \Gamma_1(s) \cdot \Gamma_2(s) \) the number of transitions of the transition function \( \delta \). At every state \( s \in S \), player 1 chooses a move \( a_1 \in \Gamma_1(s) \), and simultaneously and independently player 2 chooses a move \( a_2 \in \Gamma_2(s) \). The game then proceeds to the successor state \( t \) with probability \( \delta(s, a_1, a_2)(t) \), for all \( t \in S \). A state \( s \) is an absorbing state if for all \( a_1 \in \Gamma_1(s) \) and \( a_2 \in \Gamma_2(s) \), we have \( \delta(s, a_1, a_2)(s) = 1 \). In other words, at an absorbing state \( s \) for all choices of moves of the two players, the successor state is always \( s \).

Plays. A play \( \omega \) of \( G \) is an infinite sequence \( \omega = \langle s_0, s_1, s_2, \ldots \rangle \) of states in \( S \) such that for all \( k \geq 0 \), there are moves \( a_1^k \in \Gamma_1(s_k) \) and \( a_2^k \in \Gamma_2(s_k) \) with \( \delta(s_k, a_1^k, a_2^k)(s_{k+1}) > 0 \). We denote by \( \Omega \) the set of all plays, and by \( \Omega_s \) the set of all plays \( \omega = \langle s_0, s_1, s_2, \ldots \rangle \) such that \( s_0 = s \), that is, the set of plays starting from state \( s \).

Selectors and strategies. A selector \( \xi \) for player \( i \in \{1, 2\} \) is a function \( \xi : S \to \mathcal{D}(M) \) such that for all states \( s \in S \) and moves \( a \in M \), if \( \xi(s)(a) > 0 \), then \( a \in \Gamma_i(s) \). We denote by \( \Lambda_i \) the set of all selectors for player \( i \in \{1, 2\} \). The selector \( \xi \) is pure if for every state \( s \in S \), there is a move \( a \in M \) such that \( \xi(s)(a) = 1 \). A strategy for player \( i \in \{1, 2\} \) is a function \( \pi : S^+ \to \mathcal{D}(M) \) that associates with every finite, nonempty sequence of states, representing
the history of the play so far, a selector for player $i$; that is, for all $w \in S^+$ and $s \in S$, we have $\text{Supp}(\pi(w \cdot s)) \subseteq \Gamma_i(s)$. The strategy $\pi$ is pure if it always chooses a pure selector; that is, for all $w \in S^+$, there is a move $a \in M$ such that $\pi(w)(a) = 1$. A memoryless strategy is independent of the history of the play and depends only on the current state. Memoryless strategies correspond to selectors; we write $\xi$ for the memoryless strategy consisting in playing forever the selector $\xi$. A strategy is pure memoryless if it is both pure and memoryless. We denote by $\Pi_1$ and $\Pi_2$ the sets of all strategies for player 1 and player 2, respectively.

**Destinations of moves and selectors.** For all states $s \in S$ and moves $a_1 \in \Gamma_1(s)$ and $a_2 \in \Gamma_2(s)$, we indicate by $\text{Dest}(s, a_1, a_2) = \text{Supp}(\delta(s, a_1, a_2))$ the set of possible successors of $s$ when the moves $a_1$ and $a_2$ are chosen. Given a state $s$, and selectors $\xi_1$ and $\xi_2$ for the two players, we denote by

$$\text{Dest}(s, \xi_1, \xi_2) = \bigcup_{a_1 \in \text{Supp}(\xi_1(s)), \ a_2 \in \text{Supp}(\xi_2(s))} \text{Dest}(s, a_1, a_2)$$

the set of possible successors of $s$ with respect to the selectors $\xi_1$ and $\xi_2$.

Once a starting state $s$ and strategies $\pi_1$ and $\pi_2$ for the two players are fixed, the game is reduced to an ordinary stochastic process. Hence, the probabilities of events are uniquely defined, where an event $A \subseteq \Omega$ is a measurable set of plays. For an event $A \subseteq \Omega$, we denote by $\text{Pr}^{\pi_1, \pi_2}(A)$ the probability that a play belongs to $A$ when the game starts from $s$ and the players follow the strategies $\pi_1$ and $\pi_2$. Similarly, for a measurable function $f : \Omega \to \mathbb{R}$, we denote by $E^{\pi_1, \pi_2}(f)$ the expected value of $f$ when the game starts from $s$ and the players follow the strategies $\pi_1$ and $\pi_2$. For $i \geq 0$, we denote by $\Theta_i : \Omega \to S$ the random variable denoting the $i$-th state along a play.

**Valuations.** A valuation $v : S \to [0, 1]$ associating a real number $v(s) \in [0, 1]$ with each state $s$. Given two valuations $v, w : S \to \mathbb{R}$, we write $v \leq w$ when $v(s) \leq w(s)$ for all states $s \in S$. For an event $A$, we denote by $\text{Pr}^{\pi_1, \pi_2}(A)$ the valuation $S \to [0, 1]$ defined for all states $s \in S$ by $\{ \text{Pr}^{\pi_1, \pi_2}(A)(s) = \text{Pr}^{\pi_1, \pi_2}(A) \}$. Similarly, for a measurable function $f : \Omega \to [0, 1]$, we denote by $E^{\pi_1, \pi_2}(f)$ the valuation $S \to [0, 1]$ defined for all states $s \in S$ by $\{ E^{\pi_1, \pi_2}(f)(s) = E^{\pi_1, \pi_2}(f) \}$.

Given a valuation $v$, and two selectors $\xi_1 \in \Pi_1$ and $\xi_2 \in \Pi_2$, we define the valuations $\text{Pre}_{\xi_1, \xi_2}(v)$, $\text{Pre}_{\xi, \xi_1}(v)$, and $\text{Pre}_{\xi_1}(v)$ as follows, for all states $s \in S$:

$$\text{Pre}_{\xi_1, \xi_2}(v)(s) = \sum_{a, b \in M} \sum_{t \in S} v(t) \cdot \delta(s, a, b)(t) \cdot \xi_1(s)(a) \cdot \xi_2(s)(b)$$

$$\text{Pre}_{\xi_1}(v)(s) = \sup_{\xi_2 \in \Pi_2} \text{Pre}_{\xi_1, \xi_2}(v)(s)$$

$$\text{Pre}_{\xi}(v)(s) = \inf_{\xi_2 \in \Pi_2} \text{Pre}_{\xi_1, \xi_2}(v)(s)$$

Intuitively, $\text{Pre}_{\xi_1}(v)(s)$ is the greatest expectation of $v$ that player 1 can guarantee at a successor state of $s$. Also note that given a valuation $v$, the computation of $\text{Pre}_{\xi_1}(v)$ reduces to the solution of a zero-sum one-shot matrix game, and can be solved by linear programming. Similarly, $\text{Pre}_{\xi_1, \xi_2}(v)(s)$ is the greatest expectation of $v$ that player 1 can guarantee at a successor state of $s$ by playing the selector $\xi_1$. Note that all of these operators on valuations are monotonic: for two valuations $v, w$, if $v \leq w$, then for all selectors $\xi_1 \in \Pi_1$ and $\xi_2 \in \Pi_2$, we have $\text{Pre}_{\xi_1, \xi_2}(v) \leq \text{Pre}_{\xi_1, \xi_2}(w)$, $\text{Pre}_{\xi_1}(v) \leq \text{Pre}_{\xi_1}(w)$, and $\text{Pre}_{\xi_1}(v) \leq \text{Pre}_{\xi_1}(w)$.

**Reachability and safety objectives.** Given a subset $T \subseteq S$ of target states, the objective of a reachability game consists in reaching $T$. Therefore, we define the set winning plays as the set $\text{Reach}(T) = \{ s_0, s_1, s_2, \ldots \} \in \Omega$ for some $k \geq 0$ of plays that visit $T$. For all $s \in T$, the set $\text{Reach}(T)$ is measurable. The probability of reaching $T$ from a state $s \in S$ under strategies $\pi_1$ and $\pi_2$ for players 1 and 2, respectively, is $\text{Pr}^{\pi_1, \pi_2}(\text{Reach}(T))$. We define the value for player 1 of the reachability game with target $T$ from the state $s \in S$ as

$$\text{value}^{\pi_1}(\text{Reach}(T))(s) = \sup_{\pi_2 \in \Pi_2} \inf_{\pi_1 \in \Pi_1} \text{Pr}^{\pi_1, \pi_2}(\text{Reach}(T)).$$

Given a player-1 strategy $\pi_1$, we use the notation

$$\text{value}^{\pi_1}(\text{Reach}(T))(s) = \inf_{\pi_2 \in \Pi_2} \text{Pr}^{\pi_1, \pi_2}(\text{Reach}(T)).$$

A strategy $\pi_1$ for player 1 is optimal if for all states $s \in S$, we have

$$\text{value}^{\pi_1}(\text{Reach}(T))(s) = \text{value}^{\pi_1}(\text{Reach}(T))(s).$$

For $\epsilon > 0$, a strategy $\pi_1$ for player 1 is $\epsilon$-optimal if for all states $s \in S$, we have

$$\text{value}^{\pi_1}(\text{Reach}(T))(s) \geq \text{value}^{\pi_1}(\text{Reach}(T))(s) - \epsilon.$$
Reachability and safety objectives are dual, i.e., we have \( \text{Reach}(T) = \Omega \setminus \text{Safe}(S \setminus T) \). The quantitative determinacy result of [16] ensures that for all states \( s \in S \), we have
\[
\langle 1 \rangle(\text{Reach}(T))(s) + \langle 2 \rangle(\text{Safe}(S \setminus T))(s) = 1.
\]

### 3. Markov Decision Processes

To develop our arguments, we need some facts about one-player versions of concurrent stochastic games, known as Markov decision processes (MDPs) [11, 1]. For \( i \in \{1, 2\} \), a player-\( i \) MDP (for short, \( i \)-MDP) is a concurrent game where, for all states \( s \in S \), we have \( |\Gamma_{3-i}(s)| = 1 \). Given a concurrent game \( G \), if we fix a memoryless strategy corresponding to a selector \( \xi_i \) for player 1, the game is equivalent to a 2-MDP \( G_{\xi_i} \), with the transition function
\[
\delta_{\xi_i}(s, a_2)(t) = \sum_{a_1 \in \Gamma_i(s)} \delta(s, a_1, a_2)(t) \cdot \xi_i(s)(a_1),
\]
for all \( s \in S \) and \( a_2 \in \Gamma_2(s) \). Similarly, if we fix selectors \( \xi_1 \) and \( \xi_2 \) for both players in a concurrent game \( G \), we obtain a Markov chain, which we denote by \( G_{\xi_1, \xi_2} \).

**End components.** In an MDP, the sets of states that play an equivalent role to the closed recurrent classes of Markov chains [15] are called “end components” [6, 7].

**Definition 2 (End components)** An end component of an \( i \)-MDP \( G \), for \( i \in \{1, 2\} \), is a subset \( C \subseteq S \) of the states such that there is a selector \( \xi \) for player \( i \) so that \( C \) is a closed recurrent class of the Markov chain \( G_{\xi} \).

It is not difficult to see that an equivalent characterization of an end component \( C \) is the following. For each state \( s \in C \), there is a subset \( M_i(s) \subseteq \Gamma_i(s) \) of moves such that:

1. (closed) if a move in \( M_i(s) \) is chosen by player \( i \) at state \( s \), then all successor states that are obtained with nonzero probability lie in \( C \); and
2. (recurrent) the graph \((C, E)\), where \( E \) consists of the transitions that occur with nonzero probability when moves in \( M_i(\cdot) \) are chosen by player \( i \), is strongly connected.

Given a play \( \omega \in \Omega \), we denote by \( \text{Inf}(\omega) \) the set of states that occurs infinitely often along \( \omega \). Given a set \( \mathcal{F} \subseteq 2^S \) of subsets of states, we denote by \( \text{Inf}(\mathcal{F}) \) the event \( \{ \omega \mid \text{Inf}(\omega) \in \mathcal{F} \} \). The following theorem states that in a 2-MDP, for every strategy of player 2, the set of states that are visited infinitely often is, with probability 1, an end component. Corollary 1 follows easily from Theorem 1.

**Theorem 1** [7] For a player-1 selector \( \xi_1 \), let \( C \) be the set of end components of a 2-MDP \( G_{\xi_1} \). For all player-2 strategies \( \pi_2 \) and all states \( s \in S \), we have
\[
\Pr_{s}^{\xi_1, \pi_2}(\text{Inf}(C)) = 1.
\]

**Corollary 1** For a player-1 selector \( \xi_i \), let \( C \) be the set of end components of a 2-MDP \( G_{\xi_i} \), and let \( Z = \bigcup_{C \in \mathcal{C}} C \) be the set of states of all end components. For all player-2 strategies \( \pi_2 \) and all states \( s \in S \), we have
\[
\Pr_{s}^{\xi_i, \pi_2}(\text{Reach}(Z)) = 1.
\]

**MDPs with reachability objectives.** Given a 2-MDP with a reachability objective \( \text{Reach}(T) \) for player 2, where \( T \subseteq S \), the values can be obtained as the solution of a linear program [14]. The linear program has a variable \( x(s) \) for all states \( s \in S \), and the objective function and the constraints are as follows:
\[
\min \sum_{s \in S} x(s) \quad \text{subject to} \quad \begin{align*}
x(s) &\geq \sum_{t \in S} x(t) \cdot \delta(s, a_2)(t) \quad \text{for all } s \in S \text{ and } a_2 \in \Gamma_2(s) \\
x(s) &= 1 \quad \text{for all } s \in T \\
0 &\leq x(s) \leq 1 \quad \text{for all } s \in S
\end{align*}
\]

The correctness of the above linear program to compute the values follows from [11, 14].

### 4. Existence of Memoryless \( \varepsilon \)-Optimal Strategies for Concurrent Reachability Games

In this section we present an elementary proof of the existence of memoryless \( \varepsilon \)-optimal strategies for concurrent reachability games, for all \( \varepsilon > 0 \) (optimal strategies need not exist for concurrent games with reachability objectives [13]). A proof of the existence of memoryless optimal strategies for safety games can be found in [10].

#### 4.1. From value iteration to selectors

Consider a reachability game with target \( T \subseteq S \). Let \( W_2 = \{ s \in S \mid \langle 1 \rangle(\text{Reach}(T))(s) = 0 \} \) be the set of states from which player 1 cannot reach the target with positive probability. From [8], we know that this set can be computed as \( W_2 = \lim_{k \to \infty} W_2^k \), where \( W_0 = S \setminus T \), and for all \( k \geq 0 \),
\[
W_{2}^{k+1} = \{ s \in S \setminus T \mid \exists a_2 \in \Gamma_2(s), \forall a_1 \in \Gamma_1(s) . \text{Dest}(s, a_1, a_2) \subseteq W_2^k \}.
\]

The limit is reached in at most \( |S| \) iterations. Note that player 2 has a strategy that confines the game to \( W_2 \), and that consequently all strategies are optimal for player 1, as they realize the value 0 of the game in \( W_2 \). Therefore, without loss of generality, in the remainder we assume that all states in \( W_2 \) and \( T \) are absorbing.

Our first step towards proving the existence of memoryless \( \varepsilon \)-optimal strategies for reachability games consists in
As the strategies of memoryless sequence play in a wise fashion. Since we assume that all states in $T$ are absorbing, the classical equation reduces to the simpler equation given by (1). From the monotonicity of $P_r e_1$ it follows that $u_k \leq u_{k+1}$, that is, $P_r e_1(u_k) \geq u_k$, for all $k \geq 0$. The result of [10] establishes by a combinatorial argument that $\langle \langle P_r e_1(\text{Reach}(T)) = \lim_{k \to \infty} u_k \rangle \rangle$, where the limit is interpreted in pointwise fashion. For all $k \geq 0$, let the player-1 selector $\xi_k$ be a value-optimal selector for $u_k$, that is, a selector such that $P_r e_1(u_k) = P_r e_1(\xi_k(u_k))$. An $\varepsilon$-optimal strategy $\pi_1^k$ for player 1 can be constructed by applying the sequence $\xi_k, \xi_{k-1}, \ldots, \xi_0, \xi_0, \xi_0, \ldots$ of selectors, where the last selector, $\xi_0$, is repeated forever. It is possible to prove by induction on $k$ that

$$\inf_{\pi_2 \in \Pi_2} \mathbb{P}^{\pi_1^k, \pi_2}(\exists j \in [0..k], \Theta_j \in T) \geq u_k.$$ 

As the strategies $\pi_1^k$, for $k \geq 0$, are not necessarily memoryless, this proof does not suffice for showing the existence of memoryless $\varepsilon$-optimal strategies. On the other hand, the following example shows that the memoryless strategy $\overline{\xi}_k$ does not necessarily guarantee the value $u_k$.

**Example 1** Consider the 1-MDP shown in Fig 1. At all states except $s_3$, the set of available moves for player 1 is a singleton set. At $s_3$, the available moves for player 1 are $a$ and $b$. The transitions at the various states are shown in the figure. The objective of player 1 is to reach the state $s_0$.

We consider the value-iteration procedure and denote by $u_k$ the valuation after $k$ iterations. Writing a valuation $v$ as the list of values $(u(s_0), u(s_1), \ldots, u(s_4))$, we have:

- $u_0 = (1, 0, 0, 0, 0)$
- $u_1 = P_r e_1(u_0) = (1, 0, 1/2, 0, 0)$
- $u_2 = P_r e_1(u_1) = (1, 0, 1/2, 1/2, 0)$
- $u_3 = P_r e_1(u_2) = (1, 0, 1/2, 1/2, 1/2)$
- $u_4 = P_r e_1(u_3) = u_3 = (1, 0, 1/2, 1/2, 1/2)$

The valuation $u_3$ is thus a fixpoint.

Now consider the selector $\xi_1$ for player 1 that chooses at state $s_3$ the move $a$ with probability 1. The selector $\xi_1$ is optimal with respect to the valuation $u_3$. However, if player 1 follows the memoryless strategy $\overline{\xi}_1$, then the play visits $s_3$ and $s_4$ alternately and reaches $s_0$ with probability 0. Thus, $\xi_1$ is an example of a selector that is value-optimal, but not optimal.

![Figure 1. An MDP with reachability objective.](image)

On the other hand, consider any selector $\xi_1'$ for player 1 that chooses move $b$ at state $s_3$ with positive probability. Under the memoryless strategy $\overline{\xi}_1'$, the set \{s_0, s_1\} of states is reached with probability 1, and $s_0$ is reached with probability $1/2$. Such a $\xi_1'$ is thus an example of a selector that is both value-optimal and optimal.

In the example, the problem is that the strategy $\overline{\xi}_1$ may cause player 1 to stay forever in $S \setminus (T \cup W_2)$ with positive probability. We call “proper” the strategies of player 1 that guarantee reaching $T \cup W_2$ with probability 1.

**Definition 3 (Proper strategies and selectors)** A player-1 strategy $\pi_1$ is proper if for all player-2 strategies $\pi_2$, and for all states $s \in S \setminus (T \cup W_2)$, we have $Pr_{\pi_1, \pi_2}(\text{Reach}(T \cup W_2)) = 1$. A player-1 selector $\xi_1$ is proper if the memoryless player-1 strategy $\overline{\xi}_1$ is proper.

We note that proper strategies are closely related to Condon’s notion of a halting game [4]: precisely, a game is halting iff all player-1 strategies are proper. We can check whether a selector for player 1 is proper by considering only the pure selectors for player 2.

**Lemma 1** Given a selector $\xi_1$ for player 1, the memoryless player-1 strategy $\overline{\xi}_1$ is proper iff for every pure selector $\xi_2$ for player 2, and for all states $s \in S$, we have $Pr_{\overline{\xi}_1, \xi_2}(\text{Reach}(T \cup W_2)) = 1$.

**Proof.** We prove the contrapositive. Given a player-1 selector $\xi_1$, consider the 2-MDP $G_{\xi_1}$. If $\overline{\xi}_1$ is not proper, then by Theorem 1, there must exist an end component $C \subseteq S \setminus (T \cup W_2)$ in $G_{\xi_1}$. Then, from $C$, player 2 can avoid reaching $T \cup W_2$ by repeatedly applying a pure selector $\xi_2$ that at every state $s \in C$ deterministically chooses a move $a_2 \in G_2(s)$ such that $\text{Dest}(s, \xi_1, a_2) \subseteq C$. The existence of a suitable $\xi_2(s)$ for all states $s \in C$ follows from the definition of end component.

The following lemma shows that the selector that chooses all available moves uniformly at random is proper. This fact will be used later to initialize our strategy-improvement algorithm.

**Lemma 2** Let $\xi_{\text{unif}}^{\text{unif}}$ be the player-1 selector that at all states $s \in S \setminus (T \cup W_2)$ chooses all moves in $\Gamma_1(s)$ uniformly at random. Then $\xi_{\text{unif}}^{\text{unif}}$ is proper.

**Proof.** Assume towards contradiction that $\xi_{\text{unif}}^{\text{unif}}$ is not proper. From Theorem 1, in the 2-MDP $G_{\xi_{\text{unif}}^{\text{unif}}}$ there must
be an end component $C \subseteq S \setminus (T \cup W_2)$. Then, when player 1 follows the strategy $\bar{\xi}^\text{num}_k$, player 2 can confine the game to $C$. By the definition of $\bar{\xi}^\text{num}_k$, player 2 can ensure that the game does not leave $C$ regardless of the moves chosen by player 1, and thus, for all strategies of player 1. This contradicts the fact that $W_2$ contains all states from which player 2 can ensure that $T$ is not reached.

The following lemma shows that if the player-1 selector $\bar{\xi}_k$ computed by the value-iteration scheme (1) is proper, then the player-1 strategy $\bar{\xi}^\text{num}_k$ guarantees the value $u_k$, for all $k \geq 0$.

**Lemma 3** Let $v$ be a valuation such that $\text{Pre}^1(v) \geq v$ and $v(s) = 0$ for all states $s \in W_2$. Let $\xi_1$ be a selector for player 1 such that $\text{Pre}_{1,\xi_1}(v) = \text{Pre}^1(v)$. If $\xi_1$ is proper, then for all player-2 strategies $\pi_2$, we have $\text{Pr}_{\bar{\xi}^\text{num}_k}(\text{Reach}(T)) \geq v$.

**Proof.** Consider an arbitrary player-2 strategy $\pi_2$, and for $k \geq 0$, let $v_k = E^{\bar{\xi}_1, \pi_2}(v(\Theta_k))$ be the expected value of $v$ after $k$ steps under $\bar{\xi}_1$ and $\pi_2$. By induction on $k$, we can prove $v_k \geq v$ for all $k \geq 0$. In fact, $v_0 = v$, and for $k \geq 0$, we have

$$v_{k+1} = \text{Pre}_{1,\xi_1}(v_k) \geq \text{Pre}_{1,\xi_1}(v) = \text{Pre}^1(v) \geq v.$$  

For all $k \geq 0$ and $s \in S$, we can write $v_k$ as

$$v_k(s) = E^{\bar{\xi}_1, \pi_2}(v(\Theta_k) | \Theta_k \in T) \cdot \text{Pr}_{\bar{\xi}_1, \pi_2}(\Theta_k \in T) + \left( E^{\bar{\xi}_1, \pi_2}(v(\Theta_k) | \Theta_k \in S \setminus (T \cup W_2)) \cdot \text{Pr}_{\bar{\xi}_1, \pi_2}(\Theta_k \in S \setminus (T \cup W_2)) + E^{\bar{\xi}_1, \pi_2}(v(\Theta_k) | \Theta_k \in W_2) \cdot \text{Pr}_{\bar{\xi}_1, \pi_2}(\Theta_k \in W_2) \right).$$

Since $v(s) \leq 1$ when $s \in T$, the first term on the right-hand side is at most $\text{Pr}_{\bar{\xi}_1, \pi_2}(\Theta_k \in T)$. For the second term, we have $\lim_{k \to \infty} \text{Pr}_{\bar{\xi}_1, \pi_2}(\Theta_k \in S \setminus (T \cup W_2)) = 0$ by hypothesis, because $\text{Pr}_{\bar{\xi}_1, \pi_2}(\text{Reach}(T \cup W_2)) = 1$ and every state $s \in (T \cup W_2)$ is absorbing. Finally, the third term on the right-hand side is 0, as $v(s) = 0$ for all states $s \in W_2$. Hence, taking the limit with $k \to \infty$, we obtain

$$\text{Pr}_{\bar{\xi}_1, \pi_2}(\text{Reach}(T)) \geq \lim_{k \to \infty} v_k \geq v,$$

where the last inequality follows from $v_k \geq v$ for all $k \geq 0$. The desired result follows.

4.2. From value iteration to optimal selectors

Considering again the value-iteration scheme (1), since $\langle 1 \rangle(\text{Reach}(T)) = \lim_{k \to \infty} u_k$, for every $\varepsilon > 0$ there is a $k$ such that $u_k(s) \geq u_{k-1}(s) \geq \langle 1 \rangle(\text{Reach}(T))(s) - \varepsilon$ at all states $s \in S$. Lemma 3 indicates that, in order to construct a memoryless $\varepsilon$-optimal strategy, we need to construct from $u_{k-1}$ a player-1 selector $\bar{\xi}_1$ such that:

1. $\xi_1$ is value-optimal for $u_{k-1}$, that is, $\text{Pre}_{1,\xi_1}(u_{k-1}) = \text{Pre}_{1,\xi_1}(u_k)$; and
2. $\xi_1$ is proper.

To ensure the construction of a value-optimal, proper selector, we need some definitions. For $r > 0$, the value class

$$U_r = \{ s \in S \mid u_k(s) = r \}$$

consists of the states with value $r$ under the valuation $u_k$. Similarly, we define $U^r_{\infty} = \{ s \in S \mid u_k(s) = r \}$, for all $r \in \{r \}$.

For a state $s \in S$, let $\ell_k(s) = \min\{ j \leq k \mid u_j(s) = u_k(s) \}$ be the *entry time* of $s$ in $U^r_{\infty}$, that is, the least iteration $j$ in which the state $s$ has the same value as in iteration $k$. For $k \geq 0$, we define the player-1 selector $\eta_k$ as follows: if $\ell_k(s) > 0$, then

$$\eta_k(s) = \eta_{\ell_k(s)}(s) = \arg\sup_{\xi_1 \in \Lambda_1} \inf_{\xi_2 \in \Lambda_2} \text{Pre}_{\xi_1, \xi_2}(u_{\ell_k(s)-1});$$

otherwise, if $\ell_k(s) = 0$, then $\eta_k(s) = \eta_{\ell_k(s)}(s) = \bar{\xi}^\text{num}(s)$ (this definition is arbitrary, and it does not affect the remainder of the proof). In words, the selector $\eta_k$ is an optimal selector for $s$ at the iteration $\ell_k(s)$. It follows easily that $u_k = \text{Pre}_{1,\eta_k}(u_{k-1})$, that is, $\eta_k$ is also value-optimal for $u_{k-1}$, satisfying the first of the above conditions.

To conclude the construction, we need to prove that for $k$ sufficiently large (namely, for $k$ such that $u_k(s) > 0$ at all states $s \in S \setminus (T \cup W_2)$), the selector $\eta_k$ is proper. To this end we use Theorem 1, and show that for sufficiently large $k$ no end component of $G_{\eta_k}$ is entirely contained in $S \setminus (T \cup W_2)$. To reason about the end components of $G_{\eta_k}$, for a state $s \in S$ and a player-2 move $a_2 \in \Gamma_2(s)$, we write

$$\text{Dest}_k(s, a_2) = \bigcup_{a_1 \in \text{Supp}(\eta_k)} \text{Dest}(s, a_1, a_2)$$

for the set of possible successors of state $s$ when player 1 follows the strategy $\eta_k$, and player 2 chooses the move $a_2$.

**Lemma 4** Let $0 < r < 1$ and $k \geq 0$, and consider a state $s \in S \setminus (T \cup W_2)$ such that $s \in U_r^k$. For all moves $a_2 \in \Gamma_2(s)$, we have:

1. either $\text{Dest}_k(s, a_2) \cap U^k_r \neq \emptyset.$

---

1 In fact, the result holds for all $k$, even though our proof, for the sake of a simpler argument, does not show it.
2. or \( \text{Dest}_k(s, a_2) \subseteq U^k_r \), and there is a state \( t \in \text{Dest}_k(s, a_2) \) with \( \ell_k(t) < \ell_k(s) \).

**Proof.** For convenience, let \( m = \ell_k(s) \), and consider any move \( a_2 \in T_2(s) \).

- Consider first the case that \( \text{Dest}_k(s, a_2) \not\subseteq U^k_r \). Then, it cannot be that \( \text{Dest}_k(s, a_2) \subseteq U^k_{\leq r} \); otherwise, for all states \( t \in \text{Dest}_k(s, a_2) \), we would have \( u_k(t) \leq r \), and there would be at least one state \( t \in \text{Dest}_k(s, a_2) \) such that \( u_k(t) < r \), contradicting \( u_k(s) = r \) and \( \text{Pre}_{1:j_1}(u_k-1) = u_k \). So, it must be that \( \text{Dest}_k(s, a_2) \cap U^k_{\leq r} = \emptyset \).

- Consider now the case that \( \text{Dest}_k(s, a_2) \subseteq U^k_r \). Since \( u_m \leq u_k \), due to the monotonicity of the \( \text{Pre}_i \) operator and (1), we have that \( u_{m-1}(t) \leq r \) for all states \( t \in \text{Dest}_k(s, a_2) \). From \( r = u_k(s) = u_m(s) = \text{Pre}_{1:j_1}(u_{m-1}) \), it follows that \( u_{m-1}(t) = r \) for all states \( t \in \text{Dest}_k(s, a_2) \), implying that \( \ell_k(t) < m \) for all states \( t \in \text{Dest}_k(s, a_2) \).

The above lemma states that under \( \eta_k \), from each state \( i \in U^k_{r_1} \) with \( r > 0 \) we are guaranteed a probability bounded away from 0 of either moving to a higher-value class \( U^k_{r_2} \), or of moving to states within the value class that have a strictly lower entry time. Note that the states in the target set \( T \) are all in \( U^k_{r_1} \): they have entry time 0 in the value class for value 1. This implies that every state in \( S \setminus W \) has a probability bounded above zero of reaching \( T \) in at most \( n = |S| \) steps, so that the probability of staying forever in \( S \setminus (T \cup W) \) is 0. To prove this fact formally, we analyze the end components of \( G_{\eta_k} \) in light of Lemma 4.

**Lemma 5** For all \( k \geq 0 \), if for all states \( s \in S \setminus W \) we have \( u_k-1(s) > 0 \), then for all player-2 strategies \( \pi \), we have \( \text{Pr}_{\eta_k, \pi}(\text{Reach}(T \cup W)) = 1 \).

**Proof.** Since every state \( i \in (T \cup W) \) is absorbing, to prove this result, in view of Corollary 1, it suffices to show that no end component of \( G_{\eta_k} \) is entirely contained in \( S \setminus (T \cup W) \). Towards the contradiction, assume there is such an end component \( C \subseteq S \setminus (T \cup W) \). Then, we have \( C \subseteq U^k_{[r_1, r_2]} \) with \( C \cap U_{r_2} \neq \emptyset \), for some \( 0 < r_1 \leq r_2 \leq 1 \), where \( U^k_{[r_1, r_2]} = U^k_{r_1} \cap U^k_{[r_2]} \) is the union of the value classes for all values in the interval \( [r_1, r_2] \). Consider a state \( s \in U^k_{[r_1, r_2]} \) with minimal \( \ell_k \), that is, such that \( \ell_k(s) = \ell_k(t) \) for all other states \( t \in U^k_{[r_1, r_2]} \). From Lemma 4, it follows that for every move \( a_2 \in \Gamma_2(s) \), there is a state \( t \in \text{Dest}_k(s, a_2) \) such that (i) either \( t \in U^k_{r_2} \) and \( \ell_k(t) < \ell_k(s) \), (ii) or \( t \in U^k_{\geq r_2} \). In both cases, we obtain a contradiction.

The above lemma shows that \( \eta_k \) satisfies both requirements for optimal selectors spelt out at the beginning of Section 4.2. Hence, \( \eta_k \) guarantees the value \( u_k \). This proves the existence of memoryless \( \varepsilon \)-optimal strategies for concurrent reachability games.

**Theorem 2 (Memoryless \( \varepsilon \)-optimal strategies)** For every \( \varepsilon > 0 \), memoryless \( \varepsilon \)-optimal strategies exist for all concurrent games with reachability objectives.

**Proof.** Consider a concurrent reachability game with target \( T \subseteq S \). Since \( \lim_{k \to \infty} u_k = \langle 1 \rangle(\text{Reach}(T)) \), for every \( \varepsilon > 0 \) we can find \( k \in \mathbb{N} \) such that the following two assertions hold:

\[
\max_{s \in S} \left( \langle 1 \rangle(\text{Reach}(T))(s) - u_{k-1}(s) \right) < \varepsilon
\]

\[
\min_{s \in S \setminus W} u_{k-1}(s) > 0
\]

By construction, \( \text{Pre}^1_{1:j_1}(u_{k-1}) = \text{Pre}^1_{1}(u_{k-1}) = u_k \). Hence, from Lemma 3 and Lemma 5, for all player-2 strategies \( \pi_2 \), we have \( \text{Pr}^2_{\pi, \pi_2}(\text{Reach}(T)) \geq u_{k-1} \), leading to the result.

**5. Strategy Improvement**

In the previous section, we provided a proof of the existence of memoryless \( \varepsilon \)-optimal strategies for all \( \varepsilon > 0 \), on the basis of a value-iteration scheme. In this section we present a strategy-improvement algorithm for concurrent games with reachability objectives. The algorithm will produce a sequence of selectors \( \gamma_0, \gamma_1, \gamma_2, \ldots \) for player 1, such that:

1. for all \( i \geq 0 \), we have \( \langle 1 \rangle T^\gamma_i(\text{Reach}(T)) \leq \langle 1 \rangle T^\gamma_{i+1}(\text{Reach}(T)) \);
2. \( \lim_{i \to \infty} \langle 1 \rangle T^\gamma_i(\text{Reach}(T)) = \langle 1 \rangle(\text{Reach}(T)) \); and
3. if there is \( i \geq 0 \) such that \( \gamma_i = \gamma_{i+1} \), then \( \langle 1 \rangle T^\gamma_i(\text{Reach}(T)) = \langle 1 \rangle(\text{Reach}(T)) \).

Condition 1 guarantees that the algorithm computes a sequence of monotonically improving selectors. Condition 2 guarantees that the value guaranteed by the selectors converges to the value of the game, or equivalently, that for all \( \varepsilon > 0 \), there is a number \( i \) of iterations such that the memoryless player-1 strategy \( \gamma_i \) is \( \varepsilon \)-optimal. Condition 3 guarantees that if a selector cannot be improved, then it is optimal. Note that for concurrent reachability games, there may be no \( i \geq 0 \) such that \( \gamma_i = \gamma_{i+1} \), that is, the algorithm may fail to generate an optimal selector. This is because there are concurrent reachability games that do not admit optimal strategies, but only \( \varepsilon \)-optimal strategies for all \( \varepsilon > 0 \) [13, 9]. For **turn-based** reachability games, it can be easily seen that our algorithm terminates with an optimal selector.

We note that the value-iteration scheme of the previous section does not directly yield a strategy-improvement algorithm. In fact, the sequence of player-1 selectors \( \eta_t, \eta_t, \eta_t, \ldots \) computed in Section 4.1 may violate Condition 3: it is possible that for some \( i \geq 0 \) we have \( \eta_i = \eta_{i+1} \), but \( \eta_i \neq \eta_j \) for some \( j > i \). This is because the scheme of Section 4.1 is fundamentally a value-iteration scheme,
even though a selector is extracted from each valuation. The scheme guarantees that the valuations \( u_0, u_1, u_2, \ldots \) defined as in (1) converge, but it does not guarantee that the selectors \( \eta_0, \eta_1, \eta_2, \ldots \) improve at each iteration.

The strategy-improvement algorithm presented here shares an important connection with the proof of the existence of memoryless \( \varepsilon \)-optimal strategies presented in the previous section. Here, also, the key is to ensure that all generated selectors are proper. Again, this is ensured by modifying the selectors, at each iteration, only where they can be improved.

5.1. The strategy-improvement algorithm

Ordering of strategies. We let \( W_2 \) be as in Section 4.1, and again we assume without loss of generality that all states in \( W_2 \cup T \) are absorbing. We define a preorder \( \prec \) on the strategies for player 1 as follows: given two player 1 strategies \( \pi_1 \) and \( \pi_1' \), let \( \pi_1 \prec \pi_1' \) if the following two conditions hold: (i) \( \langle 1 \rangle^{x_1}(\text{Reach}(T)) \leq \langle 1 \rangle^{x_1'}(\text{Reach}(T)) \); and (ii) \( \langle 1 \rangle^{x_1}(\text{Reach}(T))(s) < \langle 1 \rangle^{x_1'}(\text{Reach}(T))(s) \) for some state \( s \in S \). Furthermore, we write \( \pi_1 \leq \pi_1' \) if either \( \pi_1 \prec \pi_1' \) or \( \pi_1 = \pi_1' \).

Informal description of Algorithm 1. We now present the strategy-improvement algorithm (Algorithm 1) for computing the values for all states in \( S \setminus (T \cup W_2) \). The algorithm iteratively improves player-1 strategies according to the preorder \( \prec \). The algorithm starts with the random selector \( \gamma_0 = \xi_0 \). At iteration \( i+1 \), the algorithm considers the memoryless player-1 strategy \( \pi_i \) and computes the value \( \langle 1 \rangle^{x_i}(\text{Reach}(T)) \). Observe that since \( \pi_i \) is a memoryless strategy, the composition of \( \langle 1 \rangle^{x_i}(\text{Reach}(T)) \) involves solving the 2-MDP \( G_{\pi_i} \). The valuation \( \langle 1 \rangle^{x_i}(\text{Reach}(T)) \) is named \( v_i \). For all states \( s \) such that \( \text{Pre}_1(v_i(s)) > v_i(s) \), the memoryless strategy at \( s \) is modified to a selector that is value-optimal for \( v_i \). The algorithm then proceeds to the next iteration. If \( \text{Pre}_1(v_i) = v_i \), the algorithm stops and returns the optimal memoryless strategy \( \pi_i \) for player 1. Unlike strategy-improvement algorithms for turn-based games (see [5] for a survey), Algorithm 1 is not guaranteed to terminate, because the value of a reachability game may not be rational.

5.2. Convergence

Lemma 6 Let \( \gamma_i \) and \( \gamma_{i+1} \) be the player-1 selectors obtained at iterations \( i \) and \( i+1 \) of Algorithm 1. If \( \gamma_i \) is proper, then \( \gamma_{i+1} \) is also proper.

Proof. Assume towards a contradiction that \( \gamma_i \) is proper and \( \gamma_{i+1} \) is not. Let \( \xi_2 \) be a pure selector for player 2 to witness that \( \gamma_{i+1} \) is not proper. Then there exist a subset \( C \subseteq S \setminus (T \cup W_2) \) such that \( C \) is a closed recurrent set of states in the Markov chain \( G_{\gamma_{i+1}, \xi_2} \). Let \( I \) be the nonempty set of states where the selector is modified to obtain \( \gamma_{i+1} \) from \( \gamma_i \); at all other states \( \gamma_i \) and \( \gamma_{i+1} \) agree.

Since \( \gamma_i \) and \( \gamma_{i+1} \) agree at all states other than the states in \( I \), and \( \gamma_i \) is a proper strategy, it follows that \( C \cap I \neq \emptyset \). Let \( U_i = \{ s \in S \setminus (T \cup W_2) \mid \langle 1 \rangle^{x_i}(\text{Reach}(T))(s) = v_i(s) = r \} \) be the value class with value \( r \) at iteration \( i \). For a state \( s \in U_i \) the following assertion holds: if \( \text{Dest}(s, \gamma_i, \xi_2) \subseteq U_i \), then \( \text{Dest}(\gamma_i, \xi_2, \gamma_{i+1}) \subseteq U_i \). Hence, if \( U_i \) is the greatest value class at iteration \( i \) with a nonempty intersection with the closed recurrent set \( C \). It easily follows that \( 0 < z < 1 \). Consider any state \( s \in I \), and let \( s \in U_q \). Since \( \text{Pre}_1(v_i(s)) > v_i(s) \), it follows that \( \text{Dest}(s, \gamma_{i+1}, \xi_2) \cap U_i \neq \emptyset \). Hence we must have \( z > q \), and therefore \( I \cap C \cap U_i = \emptyset \). Thus, for all states \( s \in U_i \cap C \), we have \( \gamma_i(s) = \gamma_{i+1}(s) \). Recall that \( z \) is the greatest value class at iteration \( i \) with a nonempty intersection with \( C \); hence \( U_i \subseteq U_q \). However, this gives us three statements that together form a contradiction: \( C \cap I \neq \emptyset \) (or else \( \gamma_i \) would not have been proper), \( I \cap C \cap U_i = \emptyset \), and \( C \subseteq U_i \).

Lemma 7 For all \( i \geq 0 \), the player-1 selector \( \gamma_i \) obtained at iteration \( i \) of Algorithm 1 is proper.

Proof. By Lemma 2 we have that \( \gamma_0 \) is proper. The result then follows from Lemma 6 and induction.

Lemma 8 Let \( \gamma_i \) and \( \gamma_{i+1} \) be the player-1 selectors obtained at iterations \( i \) and \( i+1 \) of Algorithm 1. Let \( I = \{ s \in S \mid \text{Pre}_1(v_i(s)) > v_i(s) \} \). Let \( v_i = \langle 1 \rangle^{x_i}(\text{Reach}(T)) \) and \( v_{i+1} = \langle 1 \rangle^{x_{i+1}}(\text{Reach}(T)) \). Then \( v_{i+1}(s) > \text{Pre}_1(v_i(s)) \) for all states \( s \in S \); and therefore \( v_i(s) \geq v_{i+1}(s) \) for all states \( s \in S \) and \( v_{i+1}(s) > v_i(s) \) for all states \( s \in I \).

Proof. Consider the valuations \( v_i \) and \( v_{i+1} \) obtained at iterations \( i \) and \( i+1 \), respectively, and let \( w_i \) be the valuation defined by \( w_i(s) = 1 - v_i(s) \) for all states \( s \in S \). Since \( \gamma_{i+1} \) is proper (by Lemma 7), it follows that the counter-optimal strategy for player 2 to minimize \( v_{i+1} \) is obtained by maximizing the probability to reach \( W_2 \). In fact, there are no end components in \( S \setminus (W_2 \cup T) \) in the 2-MDP \( G_{\gamma_{i+1}} \). Let

\[
\begin{align*}
    w_{i+1}(s) &= \begin{cases} 
        w_i(s) & \text{if } s \in S \setminus I; \\
        1 - \text{Pre}_1(v_i(s)) < w_i(s) & \text{if } s \in I.
    \end{cases}
\end{align*}
\]

In other words, \( w_{i+1} = 1 - \text{Pre}_1(v_i) \), and we also have \( w_{i+1} \leq w_i \). We now show that \( w_{i+1} \) is a feasible solution to the linear program for MDPs with the objective \( \text{Reach}(W_2) \), as described in Section 3. Since \( v_i = \langle 1 \rangle^{x_i}(\text{Reach}(T)) \), it follows that for all states \( s \in S \) and all moves \( a_2 \in \Gamma_2(s) \), we have

\[
w_i(s) \geq \sum_{t \in S} w_i(t) \cdot \delta_{\gamma_i}(s, a_2).
\]
Algorithm 1 Strategy-Improvement Algorithm

- Input: a concurrent game structure $G$ with target set $T$.

  0. Compute $W_2 = \{s \in S \mid \langle 1 \rangle(\text{Reach}(T))(s) = 0\}$.
   1. Let $\gamma_0 = \xi_{\text{inf}}^1$ and $i = 0$.
   2. Compute $v_0 = \langle 1 \rangle^{\gamma_0}(\text{Reach}(T))$.
   3. do
      1. Let $I = \{s \in S \setminus (T \cup W_2) \mid \text{Pre}_1(v_i)(s) > v_i(s)\}$.
      2. Let $\xi_1$ be a player-1 selector such that for all states $s \in I$, we have $\text{Pre}_1(\xi_1)(s) = \text{Pre}_1(v_i)(s) > v_i(s)$.
      3. The player-1 selector $\gamma_{i+1}$ is defined as follows: for each state $t \in S$, let
         \[
         \gamma_{i+1}(t) = \begin{cases} 
         \gamma_i(t) & \text{if } s \notin I; \\
         \xi_1(s) & \text{if } s \in I. 
         \end{cases}
         \]
      4. Compute $v_{i+1} = \langle 1 \rangle^{\gamma_{i+1}}(\text{Reach}(T))$.
      5. Let $i = i + 1$.
   } until $I = \emptyset$.

For all states $s \in S \setminus I$, we have $\gamma_i(s) = \gamma_{i+1}(s)$ and
\[w_{i+1}(s) = w_i(s),\] and since $w_{i+1} \leq w_i$, it follows that for all states $s \in S \setminus I$ and all moves $a_2 \in \Gamma_2(s)$, we have
\[w_{i+1}(s) \geq \sum_{t \in S} w_{i+1}(t) \cdot \delta_{\gamma_{i+1}}(s, a_2).\]

Since for $s \in I$ the selector $\gamma_{i+1}(s)$ is obtained as an optimal selector for $\text{Pre}_1(v_i)(s)$, it follows that for all states $s \in I$ and all moves $a_2 \in \Gamma_2(s)$, we have
\[w_{i+1}(s) \geq \sum_{t \in S} w_i(t) \cdot \delta_{\gamma_{i+1}}(s, a_2).\]

Since $w_{i+1} \leq w_i$, for all states $s \in I$ and all moves $a_2 \in \Gamma_2(s)$, we have
\[w_{i+1}(s) \geq \sum_{t \in S} w_{i+1}(t) \cdot \delta_{\gamma_{i+1}}(s, a_2).\]

Hence it follows that $w_{i+1}$ is a feasible solution to the linear program for MDPs with reachability objectives. Since the reachability valuation for player 2 for Reach($W_2$) is the least solution (observe that the objective function of the linear program is a minimizing function), it follows that $v_{i+1} \geq 1 - w_{i+1} = \text{Pre}_1(v_i)$. Thus we obtain $v_{i+1}(s) \geq v_i(s)$ for all states $s \in S$, and $v_{i+1}(s) > v_i(s)$ for all states $s \in I$.

**Proof.** We prove the two parts as follows.

1. The assertion that $\gamma_i \preceq \gamma_{i+1}$ follows from Lemma 8. If $\gamma_i = \gamma_{i+1}$, then $\text{Pre}_1(v_i) = v_i$, indicating that $v_i = \langle 1 \rangle(\text{Reach}(T))$. From Lemma 7 it follows that $\gamma_i$ is proper. Since $\gamma_i$ is proper by Lemma 3, we have $\langle 1 \rangle^{\gamma_i}(\text{Reach}(T)) \geq v_i = \langle 1 \rangle(\text{Reach}(T))$. It follows that $\gamma_i$ is optimal for player 1.

2. Let $v_0 = [T]$ and $u_0 = [T]$. We have $u_0 \leq v_0$. For all $k \geq 0$, by Lemma 8, we have $u_{k+1} \leq [T] \lor \text{Pre}_1(v_k)$. For all $k \geq 0$, let $u_{k+1} = [T] \lor \text{Pre}_1(u_k)$. By induction we conclude that for all $k \geq 0$, we have $u_k \leq v_k$. Moreover, $v_k \leq \langle 1 \rangle(\text{Reach}(T))$, that is, for all $k \geq 0$, we have
   \[u_k \leq v_k \leq \langle 1 \rangle(\text{Reach}(T)).\]

Since $\lim_{k \to \infty} u_k = \langle 1 \rangle(\text{Reach}(T))$, it follows that
\[\lim_{k \to \infty} \langle 1 \rangle^{\gamma_k}(\text{Reach}(T)) = \lim_{k \to \infty} v_k = \langle 1 \rangle(\text{Reach}(T)).\]

The theorem follows.

**Acknowledgments.** This research was supported in part by the NSF grants CCR-0225610 and CCR-0234690; by the NSF grant CCR-0132780 and the ARP grant SC20051123; and by the SNSF under the Indo-Swiss Joint Research Programme.

**References**