# Strategy Improvement for Concurrent Reachability and Safety Games $^{*,\dagger,\ddagger}$

Krishnendu Chatterjee[†] Luca de Alfaro[§] Thomas A. Henzinger[†‡]

[†] IST Austria (Institute of Science and Technology Austria)

[§] CE, University of California, Santa Cruz,USA

[‡] Computer and Communication Sciences, EPFL, Switzerland

{krish.chat,tah}@ist.ac.at, luca@soe.ucsc.edu

**Abstract**

We consider concurrent games played on graphs. At every round of a game, each player simultaneously and independently selects a move; the moves jointly determine the transition to a successor state. Two basic objectives are the safety objective to stay forever in a given set of states, and its dual, the reachability objective to reach a given set of states. 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, our proof is more elementary and more combinatorial. Second, we present a strategy-improvement (a.k.a. policy-iteration) algorithm for concurrent games with reachability objectives. We then present a strategy-improvement algorithm for concurrent games with safety objectives. Our algorithms yield sequences of player-1 strategies which ensure probabilities of winning that converge monotonically to the value of the game. Our result is significant because the strategy-improvement algorithm for safety games provides, for the first time, a way to approximate the value of a concurrent safety game *from below*. Previous methods could approximate the values of these games only from one direction, and as no rates of convergence are known, they did not provide a practical way to solve these games.

*Keywords.* Concurrent games; Reachability and safety objectives; Strategy improvement algorithms.

## 1 Introduction

We consider games played between two players on graphs. At every round of the game, each of the two players selects a move; the moves of the players then determine the transition to the successor state. A play of the game gives rise to a path in the graph. We consider the two basic objectives for the players: *reachability*

[*]Corresponding Author: Krishnendu Chatterjee, email: krish.chat@ist.ac.at, Address: Am Campus 1, IST Austria, Klosterneuburg, A3400, Austria. Telephone number: +43-2243-9000-3201, Fax Number: +43-2243-9000-2000.

[†]This paper is an improved version of the combined results that appeared in [3, 2]: this paper is a joint paper that combines the results of [3, 2], and presents detailed proofs of all the results.

[‡]**There was a gap in the proof of Theorem 4.3 of [2] regarding the convergence property of the improvement algorithm for safety games.** This is illustrated in Example 3. In the present version we prove all the required properties for a modified algorithm (Theorem 8). We thank anonymous reviewers for many insightful comments that helped us immensely, and warmly acknowledge their help.

and *safety*. The reachability goal asks player 1 to reach a given set of target states or, if randomization is needed to play the game, to maximize the probability of reaching the target set. The safety goal asks player 2 to ensure that a given set of safe states is never left or, if randomization is required, to minimize the probability of leaving the target set. The two objectives are dual, and the games are determined: the supremum probability with which player 1 can reach the target set is equal to one minus the supremum probability with which player 2 can confine the game to the complement of the target set [14].

These games on graphs can be divided into two classes: *turn-based* and *concurrent*. In turn-based games, only one player has a choice of moves at each state; in concurrent games, at each state both players choose a move, simultaneously and independently, from a set of available moves. For turn-based games, the solution of games with reachability and safety objectives has long been known. If each move determines a unique successor state, then the games are P-complete and can be solved in linear time in the size of the game graph. If, more generally, each move determines a probability distribution on possible successor states, then the problem of deciding whether a turn-based game can be won with probability greater than a given threshold $p \in [0, 1]$ is in NP $\cap$ co-NP [5], and the exact value of the game can be computed by a strategy-improvement algorithm [6], which works well in practice. These results all depend on the fact that in turn-based reachability and safety games, both players have optimal deterministic (i.e., no randomization is required), memoryless strategies. These strategies are functions from states to moves, so they are finite in number, and this guarantees the termination of the strategy-improvement algorithm.

The situation is very different for concurrent games. The player-1 *value* of the game is defined, as usual, as the sup-inf value: the supremum, over all strategies of player 1, of the infimum, over all strategies of player 2, of the probability of achieving the reachability or safety goal. In concurrent reachability games, player 1 is guaranteed only the existence of $\varepsilon$-optimal strategies, which ensure that the value of the game is achieved within a specified tolerance $\varepsilon > 0$ [14]. Moreover, while these strategies (which depend on $\varepsilon$) are memoryless, in general they require randomization [14] (even in the special case in which the transition function is deterministic). For player 2 (the safety player), *optimal* memoryless strategies exist [24], which again require randomization (even when the transition function is deterministic). All of these strategies are functions from states to probability distributions on moves. The question of deciding whether a concurrent game can be won with probability greater than $p$ is in PSPACE; this is shown by reduction to the theory of the real-closed fields [13].

To summarize: while strategy-improvement algorithms are available for turn-based reachability and safety games [6], so far no strategy-improvement algorithms or even approximation schemes were known for concurrent games. If one wanted to compute the value of a concurrent game within a specified tolerance $\varepsilon > 0$, one was reduced to using a binary search algorithm that approximates the value by iterating queries in the theory of the real-closed fields. Value-iteration schemes were known for such games, but they can be used to approximate the value from one direction only, for reachability goals from below, and for safety goals from above [11]. The value-iteration schemes are not guaranteed to terminate. Worse, since no convergence rates are known for these schemes, they provide no termination criteria for approximating a value within $\varepsilon$.

**Our results for concurrent reachability games.** Concurrent reachability games belong to the family of stochastic games [26, 14], and they have been studied more specifically in [10, 9, 11]. Our contributions for concurrent reachability games 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, using the proof techniques we developed for proving existence of memoryless $\varepsilon$-optimal strategies, for $\varepsilon > 0$, we obtain a strategy-improvement (a.k.a. policy-iteration) algorithm for concurrent reachability games. Unlike in the special case of turn-based games the algorithm need not terminate in finitely many iterations.

It has long been known that optimal strategies need not exist for concurrent reachability games, and for all $\varepsilon > 0$, there exist $\varepsilon$-optimal strategies that are memoryless [14]. A proof of this fact can be obtained by considering limit of discounted games. The proof considers *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 $0 < \alpha \leq 1$. It is possible to show that, for $0 < \alpha < 1$, memoryless optimal strategies 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 from the field of real Puisieux series [23]. This proof idea works not only for reachability games, but also for total-reward games with nonnegative rewards (see [15, Chapter 5] for details). A more recent result [13] establishes the existence of memoryless $\varepsilon$-optimal strategies for certain infinite-state (recursive) concurrent games, but again the proof relies on results from analysis and properties of solutions of certain polynomial functions. Another proof of existence of memoryless $\varepsilon$-optimal strategies for reachability objectives follows from the result of [14] and the proof uses induction on the number of states of the game. We show the existence of memoryless $\varepsilon$-optimal strategies for concurrent reachability games by more combinatorial and elementary means. Our proof relies only on combinatorial techniques and on simple properties of Markov decision processes [1, 8]. As our proof is more combinatorial, we believe that the proof techniques 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 [11]. 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 $s$. The convergence is monotonic from below, but no rate of convergence was known. From each valuation $u_i$, we can 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 play 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 $\varepsilon > 0$. Our proof is completely different as compared to the proof of [14]: the proof of [14] uses induction on the number of states, whereas our proof is based on the notion of ranking function obtained from the value-iteration algorithm.

We then apply the techniques developed for the above proof to design a *strategy-improvement* algorithm for concurrent reachability games. Strategy-improvement algorithms, also known as *policy-iteration* algorithms in the context of Markov decision processes [20], 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 generated strategies are proper. Unlike previous proofs of strategy-improvement algorithms for concurrent games [6, 15], which rely on the analysis of discounted versions of the games, our analysis is again more combinatorial. Hoffman-Karp [19] presented a strategy improvement algorithm for the special case of concurrent games with ergodic property (i.e., from every state $s$ any other state $t$ can be guaranteed to reach with probability 1) (also see algorithm for discounted games in [25]). Observe that for concurrent reachability games, with the ergodic assumption the value at all states is trivially 1, and thus the ergodic assumption gives us the trivial case. Our results give a combinatorial strategy improvement algorithm for the whole class of concurrent reachability games. The results of [13] presents a strategy improvement algorithm for recursive concurrent games with termination criteria: the algorithm of [13] is more involved (depends on properties of certain polynomial functions) and works for the more general class of recursive concurrent games. Differently from turn-based games [6], for concurrent games we cannot guarantee the termination of the strategy-improvement algorithm. However, for turn-based stochastic games we present a detailed analysis of termination criteria. Our analysis is based on bounds on the precision of values for turn-based stochastic games. As a consequence of our analysis, we obtain an improved upper bound for termination for turn-based stochastic games.

**Our results for concurrent safety games.** We present for the first time a strategy-improvement scheme that approximates the value of a concurrent safety game *from below*. Together with the strategy improvement algorithm for reachability games, or the value-iteration scheme, to approximate the value of such a game from above, we obtain a termination criterion for computing the value of concurrent reachability and safety games within any given tolerance $\varepsilon > 0$. This is the first termination criterion for an algorithm that approximates the value of a concurrent game. Several difficulties had to be overcome in developing our scheme. First, while the strategy-improvement algorithm that approximates reachability values from below is based on locally improving a strategy on the basis of the valuation it yields, this approach does not suffice for approximating safety values from below: we would obtain an increasing sequence of values, but they would not necessarily converge to the value of the game (see Example 2). Rather, we introduce a novel, non-local improvement step, which augments the standard valuation-based improvement step. Each non-local step involves the solution of an appropriately constructed turn-based game. The turn-based game constructed is polynomial in the state space of the original game, but *exponential* in the number of actions. It is an interesting open question whether the turn-based game can be also made polynomial in the number of the actions. Second, as value-iteration for safety objectives converges from above, while our sequences of strategies yield values that converge from below, the proof of convergence for our algorithm cannot be derived from a connection with value-iteration, as was the case for reachability objectives. We had to develop new proof techniques both to show the monotonicity of the strategy values produced by our algorithm, and to show their convergence to the value of the game.

**Added value of our algorithms.** The new strategy improvement algorithms we present in this paper has two important contributions as compared to the classical value-iteration algorithms.

1. *Termination for approximation.* The value-iteration algorithm for reachability games converges from below, and the value-iteration for safety games converges for above. Hence given desired precision $\varepsilon > 0$ for approximation, there is no termination criteria to stop the value-iteration algorithm and guarantee $\varepsilon$-approximation. The sequence of valuation of our strategy improvement algorithm for concurrent safety games converges from below, and along with the value-iteration or strategy improvement algorithm for concurrent reachability games we obtain the *first* termination criteria for $\varepsilon$-approximation of values in concurrent reachability and safety games. Using a result of [18] on the bound on $k$-uniform memoryless $\varepsilon$-optimal strategies, for $\varepsilon > 0$, we also obtain a bound on the number of iterations of the strategy improvement algorithms that guarantee $\varepsilon$-approximation of the values. Moreover a recent result of [17] provide a nearly tight double exponential upper and lower bound on the number of iterations required for $\varepsilon$-approximation of the values.

2. *Approximation of strategies.* Our strategy improvement algorithms are also the first approach to approximate memoryless $\varepsilon$-optimal strategies in concurrent reachability and safety games. The witness

strategy produced by the value-iteration algorithm for concurrent reachability games is not memoryless; and for concurrent safety games since the value-iteration algorithm converges from above it does not provide any witness strategies. Our strategy improvement algorithms for concurrent reachability and safety games yield sequence of memoryless strategies that ensure for convergence to the value of the game from below, and yield witness memoryless strategies to approximate the value of concurrent reachability and safety games.

# 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 $Supp(\delta) = \{x \in A \mid \delta(x) > 0\}$ 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 or actions.
- Two move assignments $\Gamma_1, \Gamma_2 \colon S \to 2^M \setminus \emptyset$. 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 $|\delta|$ the size of transition function, i.e., $|\delta| = \sum_{s \in S, a \in \Gamma_1(s), b \in \Gamma_2(s), t \in S} |\delta(s,a,b)(t)|$, where $|\delta(s,a,b)(t)|$ is the number of bits required to specify the transition probability $\delta(s,a,b)(t)$. We denote by $|G|$ the size of the game graph, and $|G| = |\delta| + |S|$. 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 $s$.

**Definition 2** (TURN-BASED STOCHASTIC GAMES). A *turn-based stochastic game graph* ($2^1/_2$*-player game graph*) $G = \langle (S,E), (S_1, S_2, S_R), \delta \rangle$ consists of a finite directed graph $(S,E)$, a partition ($S_1$, $S_2$, $S_R$) of the finite set $S$ of states, and a probabilistic transition function $\delta$: $S_R \to \mathcal{D}(S)$, where $\mathcal{D}(S)$ denotes the set of probability distributions over the state space $S$. The states in $S_1$ are the *player*-1 states, where player 1 decides the successor state; the states in $S_2$ are the *player*-2 states, where player 2 decides the successor state; and the states in $S_R$ are the *random or probabilistic* states, where the successor state is chosen according to the probabilistic transition function $\delta$. We assume that for $s \in S_R$ and $t \in S$, we have $(s,t) \in E$ iff $\delta(s)(t) > 0$, and we often write $\delta(s,t)$ for $\delta(s)(t)$. For technical convenience we assume that every state in the graph $(S,E)$ has at least one outgoing edge. For a state $s \in S$, we write $E(s)$ to denote the set $\{t \in S \mid (s,t) \in E\}$ of possible successors. We denote by $|\delta|$ the size of the transition function, i.e., $|\delta| = \sum_{s \in S_R, t \in S} |\delta(s)(t)|$, where $|\delta(s)(t)|$ is the number of bits required to specify the transition probability $\delta(s)(t)$. We denote by $|G|$ the size of the game graph, and $|G| = |\delta| + |S| + |E|$.

**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 $a \in \Gamma_i(s)$. A selector $\xi$ for player $i$ at a state $s$ is a distribution over moves such that if $\xi(s)(a) > 0$, then $a \in \Gamma_i(s)$. We denote by $\Lambda_i$ the set of all selectors for player $i \in \{1, 2\}$, and similarly, we denote by $\Lambda_i(s)$ the set of all selectors for player $i$ at a state $s$. 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 $Supp(\pi(w \cdot s)) \subseteq \Gamma_i(s)$. 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 $\overline{\xi}$ for the memoryless strategy consisting in playing forever the selector $\xi$. A strategy is *pure memoryless* if it is both pure and memoryless. In a turn-based stochastic game, a strategy for player 1 is a function $\pi_1 : S^* \cdot S_1 \to \mathcal{D}(S)$, such that for all $w \in S^*$ and for all $s \in S_1$ we have $Supp(\pi_1(w \cdot s)) \subseteq E(s)$. Memoryless strategies and pure memoryless strategies are obtained as the restriction of strategies as in the case of concurrent game graphs. The family of strategies for player 2 are defined analogously. We denote by $\Pi_1$ and $\Pi_2$ the sets of all strategies for player 1 and player 2, respectively. We denote by $\Pi_i^M$ and $\Pi_i^{PM}$ the sets of memoryless strategies and pure memoryless strategies for player $i$, 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 $Dest(s, a_1, a_2) = Supp(\delta(s, a_1, a_2))$ 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

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

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* $\mathcal{A} \subseteq \Omega_s$ is a measurable set of plays. For an event $\mathcal{A} \subseteq \Omega_s$, we denote by $\Pr_s^{\pi_1,\pi_2}(\mathcal{A})$ the probability that a play belongs to $\mathcal{A}$ when the game starts from $s$ and the players follows the strategies $\pi_1$ and $\pi_2$. Similarly, for a measurable function $f : \Omega_s \to \mathbb{R}$, we denote by $\mathrm{E}_s^{\pi_1,\pi_2}(f)$ 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* is a mapping $v : S \to [0, 1]$ 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 $\mathcal{A}$, we denote by $\Pr^{\pi_1,\pi_2}(\mathcal{A})$ the valuation $S \to [0, 1]$ defined for all states $s \in S$ by $\big(\Pr^{\pi_1,\pi_2}(\mathcal{A})\big)(s) = \Pr_s^{\pi_1,\pi_2}(\mathcal{A})$. Similarly, for a measurable function $f : \Omega_s \to [0, 1]$, we denote by $\mathrm{E}^{\pi_1,\pi_2}(f)$ the valuation $S \to [0, 1]$ defined for all $s \in S$ by $\big(\mathrm{E}^{\pi_1,\pi_2}(f)\big)(s) = \mathrm{E}_s^{\pi_1,\pi_2}(f)$.

**The $Pre$ operator.** Given a valuation $v$, and two selectors $\xi_1 \in \Lambda_1$ and $\xi_2 \in \Lambda_2$, we define the valuations $Pre_{\xi_1,\xi_2}(v)$, $Pre_{1:\xi_1}(v)$, and $Pre_1(v)$ as follows, for all states $s \in S$:

$$Pre_{\xi_1,\xi_2}(v)(s) = \sum_{a,b \in M} \sum_{t \in S} v(t) \cdot \delta(s,a,b)(t) \cdot \xi_1(s)(a) \cdot \xi_2(s)(b)$$
$$Pre_{1:\xi_1}(v)(s) = \inf_{\xi_2 \in \Lambda_2} Pre_{\xi_1,\xi_2}(v)(s)$$
$$Pre_1(v)(s) = \sup_{\xi_1 \in \Lambda_1} \inf_{\xi_2 \in \Lambda_2} Pre_{\xi_1,\xi_2}(v)(s)$$

Intuitively, $Pre_1(v)(s)$ 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 $Pre_1(v)$ reduces to the solution of a zero-sum one-shot matrix game, and can be solved by linear programming. Similarly, $Pre_{1:\xi_1}(v)(s)$ 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 $\xi_1 \in \Lambda_1$ and $\xi_2 \in \Lambda_2$, we have $Pre_{\xi_1,\xi_2}(v) \leq Pre_{\xi_1,\xi_2}(w)$, $Pre_{1:\xi_1}(v) \leq Pre_{1:\xi_1}(w)$, and $Pre_1(v) \leq Pre_1(w)$.

**Reachability and safety objectives.** Given a set $F \subseteq S$ of *safe* states, the objective of a safety game consists in never leaving $F$. Therefore, we define the set of winning plays as the set $\text{Safe}(F) = \{\langle s_0, s_1, s_2, \ldots \rangle \in \Omega \mid s_k \in F \text{ for all } k \geq 0\}$. Given a subset $T \subseteq S$ of *target* states, the objective of a reachability game consists in reaching $T$. Correspondingly, the set winning plays is $\text{Reach}(T) = \{\langle s_0, s_1, s_2, \ldots \rangle \in \Omega \mid s_k \in T \text{ for some } k \geq 0\}$ of plays that visit $T$. For all $F \subseteq S$ and $T \subseteq S$, the sets $\text{Safe}(F)$ and $\text{Reach}(T)$ is measurable. An objective in general is a measurable set, and in this paper we consider only reachability and safety objectives. For an objective $\Phi$, the probability of satisfying $\Phi$ from a state $s \in S$ under strategies $\pi_1$ and $\pi_2$ for players 1 and 2, respectively, is $\Pr_s^{\pi_1,\pi_2}(\Phi)$. We define the *value* for player 1 of game with objective $\Phi$ from the state $s \in S$ as

$$\langle\langle 1 \rangle\rangle_{\mathsf{val}}(\Phi)(s) = \sup_{\pi_1 \in \Pi_1} \inf_{\pi_2 \in \Pi_2} \Pr_s^{\pi_1,\pi_2}(\Phi);$$

i.e., the value is the maximal probability with which player 1 can guarantee the satisfaction of $\Phi$ against all player 2 strategies. Given a player-1 strategy $\pi_1$, we use the notation

$$\langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\pi_1}(\Phi)(s) = \inf_{\pi_2 \in \Pi_2} \Pr_s^{\pi_1,\pi_2}(\Phi).$$

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

$$\langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\pi_1}(\Phi)(s) = \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\Phi)(s).$$

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

$$\langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\pi_1}(\Phi)(s) \geq \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\Phi)(s) - \varepsilon.$$

The notion of values and optimal strategies for player 2 are defined analogously. 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 [14] ensures that for all states $s \in S$, we have

$$\langle\langle 1 \rangle\rangle_{\mathsf{val}}(\text{Safe}(F))(s) + \langle\langle 2 \rangle\rangle_{\mathsf{val}}(\text{Reach}(S \setminus F))(s) = 1.$$

## 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) [12, 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 selector $\xi_1$ for player 1, the game is equivalent to a 2-MDP $G_{\xi_1}$ with the transition function

$$\delta_{\xi_1}(s, a_2)(t) = \sum_{a_1 \in \Gamma_1(s)} \delta(s, a_1, a_2)(t) \cdot \xi_1(s)(a_1),$$

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 [21, Chapter 4] are called "end components" [7, 8].

**Definition 3** (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 $\mathrm{Inf}(\omega)$ 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 $\mathrm{Inf}(\mathcal{F})$ the event $\{\omega \mid \mathrm{Inf}(\omega) \in \mathcal{F}\}$. 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** ([8]). *For a player-1 selector $\xi_1$, let $\mathcal{C}$ 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* $\mathrm{Pr}_s^{\overline{\xi}_1,\pi_2}(\mathrm{Inf}(\mathcal{C})) = 1$.

**Corollary 1** *For a player-1 selector $\xi_1$, let $\mathcal{C}$ be the set of end components of a 2-MDP $G_{\xi_1}$, 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* $\mathrm{Pr}_s^{\overline{\xi}_1,\pi_2}(\mathit{Reach}(Z)) = 1$.

**MDPs with reachability objectives.** Given a 2-MDP with a reachability objective $\mathrm{Reach}(T)$ for player 2, where $T \subseteq S$, the values can be obtained as the solution of a linear program [15] (see Section 2.9 of [15] where linear program solution is given for MDPs with limit-average objectives and reachability objective is a special case of limit-average objectives). The linear program has a variable $x(s)$ for all states $s \in S$, and the objective function and the constraints are as follows:

$$\min \sum_{s \in S} x(s) \quad \text{subject to}$$

$$x(s) \geq \sum_{t \in S} x(t) \cdot \delta(s, a_2)(t) \text{ for all } s \in S \text{ and } a_2 \in \Gamma_2(s)$$

$$x(s) = 1 \text{ for all } s \in T$$

$$0 \leq x(s) \leq 1 \text{ for all } s \in S$$

The correctness of the above linear program to compute the values follows from [15] (see section 2.9 of [15], and also see [7] for the correctness of the linear program).

# 4 Existence of Memoryless $\varepsilon$-Optimal Strategies for Concurrent Reachability Games

In this section we present an elementary and combinatorial 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 [14]).

## 4.1 From value iteration to selectors

Consider a reachability game with target $T \subseteq S$, i.e., objective for player 1 is $\text{Reach}(T)$. Let $W_2 = \{s \in S \mid \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\text{Reach}(T))(s) = 0\}$ be the set of states from which player 1 cannot reach the target with positive probability. From [9], we know that this set can be computed as $W_2 = \lim_{k \to \infty} W_2^k$, where $W_2^0 = S \setminus T$, and for all $k \geq 0$,

$$W_2^{k+1} = \{s \in S \setminus T \mid \exists a_2 \in \Gamma_2(s) \,.\, \forall a_1 \in \Gamma_1(s) \,.\, Dest(s, a_1, a_2) \subseteq W_2^k\} \,.$$

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 considering a value-iteration scheme for the computation of $\langle\langle 1 \rangle\rangle_{\mathsf{val}}(\text{Reach}(T))$. Let $[T] : S \to [0, 1]$ be the indicator function of $T$, defined by $[T](s) = 1$ for $s \in T$, and $[T](s) = 0$ for $s \notin T$. Let $u_0 = [T]$, and for all $k \geq 0$, let

$$u_{k+1} = Pre_1(u_k). \tag{1}$$

Note that the classical equation assigns $u_{k+1} = [T] \vee Pre_1(u_k)$, where $\vee$ is interpreted as the maximum in pointwise 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 $Pre_1$ it follows that $u_k \leq u_{k+1}$, that is, $Pre_1(u_k) \geq u_k$, for all $k \geq 0$. The result of [11] establishes by a combinatorial argument that $\langle\langle 1 \rangle\rangle_{\mathsf{val}}(\text{Reach}(T)) = \lim_{k \to \infty} u_k$, where the limit is interpreted in pointwise fashion. For all $k \geq 0$, let the player-1 selector $\zeta_k$ be a *value-optimal* selector for $u_k$, that is, a selector such that $Pre_1(u_k) = Pre_{1:\zeta_k}(u_k)$. An $\varepsilon$-optimal strategy $\pi_1^k$ for player 1 can be constructed by applying the sequence $\zeta_k, \zeta_{k-1}, \ldots, \zeta_1, \zeta_0, \zeta_0, \zeta_0, \ldots$ of selectors, where the last selector, $\zeta_0$, is repeated forever. It is possible to prove by induction on $k$ that

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

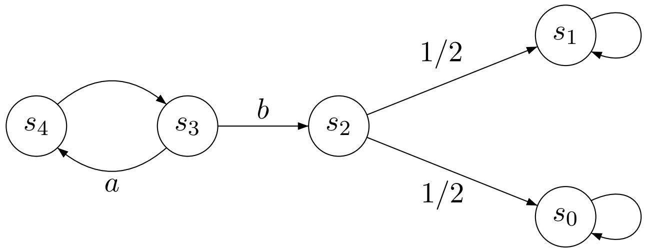


Figure 1: An MDP with reachability objective.

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 $\overline{\zeta}_k$ 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 $u$ as the list of values $\big(u(s_0), u(s_1), \ldots, u(s_4)\big)$, we have:

$$\begin{aligned} u_0 &= (1, 0, 0, 0, 0) \\ u_1 = Pre_1(u_0) &= (1, 0, 1/2, 0, 0) \\ u_2 = Pre_1(u_1) &= (1, 0, 1/2, 1/2, 0) \\ u_3 = Pre_1(u_2) &= (1, 0, 1/2, 1/2, 1/2) \\ u_4 = Pre_1(u_3) = u_3 &= (1, 0, 1/2, 1/2, 1/2) \end{aligned}$$

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.

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 4** (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 $\Pr_s^{\pi_1,\pi_2}(\text{Reach}(T \cup W_2)) = 1$. 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* [5]: 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 $\Pr_s^{\overline{\xi}_1,\overline{\xi}_2}(Reach(T \cup W_2)) = 1$.*

**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 $a_2 \in \Gamma_2(s)$ such that $Dest(s, \xi_1, a_2) \subseteq C$. 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 $\xi_1^{unif}$ 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 $\xi_1^{unif}$ is proper.*

**Proof.** Assume towards contradiction that $\xi_1^{unif}$ is not proper. From Theorem 1, in the 2-MDP $G_{\xi_1^{unif}}$ there must be an end component $C \subseteq S \setminus (T \cup W_2)$. Then, when player 1 follows the strategy $\overline{\xi}_1^{unif}$, player 2 can confine the game to $C$. By the definition of $\xi_1^{unif}$, 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 $\zeta_k$ computed by the value-iteration scheme (1) is proper, then the player-1 strategy $\overline{\zeta}_k$ guarantees the value $u_k$, for all $k \geq 0$.

**Lemma 3** *Let $v$ be a valuation such that $Pre_1(v) \geq v$ and $v(s) = 0$ for all states $s \in W_2$. Let $\xi_1$ be a selector for player 1 such that $Pre_{1:\xi_1}(v) = Pre_1(v)$. If $\xi_1$ is proper, then for all player-2 strategies $\pi_2$, we have $\Pr^{\overline{\xi}_1,\pi_2}(Reach(T)) \geq v$.*

**Proof.** Consider an arbitrary player-2 strategy $\pi_2$, and for $k \geq 0$, let

$$v_k = \mathrm{E}^{\overline{\xi}_1,\pi_2}\big(v(\Theta_k)\big)$$

be the expected value of $v$ after $k$ steps under $\overline{\xi}_1$ 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

$$v_{k+1} \geq Pre_{1:\xi_1}(v_k) \geq Pre_{1:\xi_1}(v) = Pre_1(v) \geq v.$$

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

$$\begin{aligned} v_k(s) = {} & \mathrm{E}_s^{\overline{\xi}_1,\pi_2}\big(v(\Theta_k) \mid \Theta_k \in T\big) \cdot \Pr_s^{\overline{\xi}_1,\pi_2}\big(\Theta_k \in T\big) \\ & + \mathrm{E}_s^{\overline{\xi}_1,\pi_2}\big(v(\Theta_k) \mid \Theta_k \in S \setminus (T \cup W_2)\big) \cdot \Pr_s^{\overline{\xi}_1,\pi_2}\big(\Theta_k \in S \setminus (T \cup W_2)\big) \\ & + \mathrm{E}_s^{\overline{\xi}_1,\pi_2}\big(v(\Theta_k) \mid \Theta_k \in W_2\big) \cdot \Pr_s^{\overline{\xi}_1,\pi_2}\big(\Theta_k \in W_2\big). \end{aligned}$$

Since $v(s) \leq 1$ when $s \in T$, the first term on the right-hand side is at most $\Pr_s^{\overline{\xi}_1,\pi_2}\big(\Theta_k \in T\big)$. For the second term, we have $\lim_{k\to\infty} \Pr^{\overline{\xi}_1,\pi_2}\big(\Theta_k \in S \setminus (T \cup W_2)\big) = 0$ by hypothesis, because $\Pr^{\overline{\xi}_1,\pi_2}(\mathrm{Reach}(T \cup W_2)) =$

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

$$\Pr^{\overline{\xi}_1,\pi_2}\big(\text{Reach}(T)\big) = \lim_{k\to\infty} \Pr^{\overline{\xi}_1,\pi_2}\big(\Theta_k \in T\big) \geq \lim_{k\to\infty} v_k \geq v,$$

where the last inequality follows from $v_k \geq v$ for all $k \geq 0$. Note that $v_k = \Pr^{\overline{\xi}_1,\pi_2}\big(\Theta_k \in T\big)$, and since $T$ is absorbing it follows that $v_k$ is non-deccreasing (monotonic) and is bounded by 1 (since it is a probability measure). Hence the limit of $v_k$ is defined. The desired result follows. ∎

### 4.2 From value iteration to optimal selectors

In this section we show how to obtain memoryless $\varepsilon$-optimal strategies from the value-iteration scheme, for $\varepsilon > 0$. In the following section the existence such strategies would be established using a strategy-iteration scheme. The strategy-iteration scheme has been used previously to establish existence of memoryless $\varepsilon$-optimal strategies, for $\varepsilon > 0$ (for example see [13] and also results of Condon [5] for turn-based games). However our proof which constructs the memoryless strategies based on value-iteration scheme is new. Considering again the value-iteration scheme (1), since $\langle\langle 1\rangle\rangle_{\mathsf{val}}(\text{Reach}(T)) = \lim_{k\to\infty} u_k$, for every $\varepsilon > 0$ there is a $k$ such that $u_k(s) \geq u_{k-1}(s) \geq \langle\langle 1\rangle\rangle_{\mathsf{val}}(\text{Reach}(T))(s) - \varepsilon$ 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 $\xi_1$ such that:

1. $\xi_1$ is value-optimal for $u_{k-1}$, that is, $Pre_{1:\xi_1}(u_{k-1}) = Pre_1(u_{k-1}) = u_k$; 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*

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

consists of the states with value $r$ under the valuation $u_k$. Similarly we define $U_{\bowtie r}^k = \{s \in S \mid u_k(s) \bowtie r\}$, for $\bowtie \in \{<, \leq, \geq, >\}$. 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 $U_{u_k(s)}^k$, 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

$$\eta_k(s) = \eta_{\ell_k(s)}(s) = \arg\max_{\xi_1\in\Lambda_1} \inf_{\xi_2\in\Lambda_2} Pre_{\xi_1,\xi_2}(u_{\ell_k(s)-1});$$

otherwise, if $\ell_k(s) = 0$, then $\eta_k(s) = \eta_{\ell_k(s)}(s) = \xi_1^{\mathit{unif}}(s)$ (this definition is arbitrary, and it does not affect the remainder of the proof). In words, the selector $\eta_k(s)$ is an optimal selector for $s$ at the iteration $\ell_k(s)$. It follows easily that $u_k = Pre_{1:\eta_k}(u_{k-1})$, 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 $S \setminus (T \cup W_2)$.[1] 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

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

[1] In fact, the result holds for all $k$, even though our proof, for the sake of a simpler argument, does not show it.

for the set of possible successors of state $s$ when player 1 follows the strategy $\overline{\eta}_k$, and player 2 chooses the move $a_2$.

**Lemma 4** *Let $0 < r \leq 1$ 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 $Dest_k(s, a_2) \cap U_{>r}^k \neq \emptyset$,*
2. *or $Dest_k(s, a_2) \subseteq U_r^k$, and there is a state $t \in Dest_k(s, a_2)$ with $\ell_k(t) < \ell_k(s)$.*

**Proof.** For convenience, let $m = \ell_k(s)$, and consider any move $a_2 \in \Gamma_2(s)$.

- Consider first the case that $Dest_k(s, a_2) \not\subseteq U_r^k$. Then, it cannot be that $Dest_k(s, a_2) \subseteq U_{\leq r}^k$; otherwise, for all states $t \in Dest_k(s, a_2)$, we would have $u_k(t) \leq r$, and there would be at least one state $t \in Dest_k(s, a_2)$ such that $u_k(t) < r$, contradicting $u_k(s) = r$ and $Pre_{1:\eta_k}(u_{k-1}) = u_k$. So, it must be that $Dest_k(s, a_2) \cap U_{>r}^k \neq \emptyset$.
- Consider now the case that $Dest_k(s, a_2) \subseteq U_r^k$. Since $u_m \leq u_k$, due to the monotonicity of the $Pre_1$ operator and (1), we have that $u_{m-1}(t) \leq r$ for all states $t \in Dest_k(s, a_2)$. From $r = u_k(s) = u_m(s) = Pre_{1:\eta_k}(u_{m-1})$, it follows that $u_{m-1}(t) = r$ for all states $t \in Dest_k(s, a_2)$, implying that $\ell_k(t) < m$ for all states $t \in Dest_k(s, a_2)$. ∎

The above lemma states that under $\eta_k$, from each state $i \in U_r^k$ with $r > 0$ we are guaranteed a probability bounded away from 0 of either moving to a higher-value class $U_{>r}^k$, 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 $U_1^0$: they have entry-time 0 in the value class for value 1. This implies that every state in $S \setminus W_2$ has a probability bounded above zero of reaching $T$ in at most $n = |S|$ steps, so that the probability of staying forever in $S \setminus (T \cup W_2)$ 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 $s \in S \setminus W_2$ we have $u_{k-1}(s) > 0$, then for all player-2 strategies $\pi_2$, we have $\mathrm{Pr}^{\overline{\eta}_k, \pi_2}(Reach(T \cup W_2)) = 1$.*

**Proof.** Since every state $s \in (T \cup W_2)$ 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 $S \setminus (T \cup W_2)$. Towards the contradiction, assume there is such an end component $C \subseteq S \setminus (T \cup W_2)$. Then, we have $C \subseteq U_{[r_1, r_2]}^k$ with $C \cap U_{r_2} \neq \emptyset$, for some $0 < r_1 \leq r_2 \leq 1$, where $U_{[r_1, r_2]}^k = U_{\geq r_1}^k \cap U_{\leq r_2}^k$ is the union of the value classes for all values in the interval $[r_1, r_2]$. Consider a state $s \in U_{r_2}^k$ with minimal $\ell_k$, that is, such that $\ell_k(s) \leq \ell_k(t)$ for all other states $t \in U_{r_2}^k$. From Lemma 4, it follows that for every move $a_2 \in \Gamma_2(s)$, there is a state $t \in Dest_k(s, a_2)$ such that (i) either $t \in U_{r_2}^k$ and $\ell_k(t) < \ell_k(s)$, (ii) or $t \in U_{>r_2}^k$. 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 $\lim_{k\to\infty} u_k = \langle\!\langle 1\rangle\!\rangle_{\mathsf{val}}(\text{Reach}(T))$, for every $\varepsilon > 0$ we can find $k \in \mathbb{N}$ such that the following two assertions hold:

$$\max_{s\in S}\big(\langle\!\langle 1\rangle\!\rangle_{\mathsf{val}}(\text{Reach}(T))(s) - u_{k-1}(s)\big) < \varepsilon$$

$$\min_{s\in S\setminus W_2} u_{k-1}(s) > 0$$

By construction, $Pre_{1:\eta_k}(u_{k-1}) = Pre_1(u_{k-1}) = u_k$. Hence, from Lemma 3 and Lemma 5, for all player-2 strategies $\pi_2$, we have $\Pr^{\overline{\eta}_k,\pi_2}(\text{Reach}(T)) \geq u_{k-1}$, leading to the result. ∎

# 5 Strategy Improvement Algorithm for Concurrent Reachability Games

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 $\langle\!\langle 1\rangle\!\rangle^{\overline{\gamma}_i}_{\mathsf{val}}(\text{Reach}(T)) \leq \langle\!\langle 1\rangle\!\rangle^{\overline{\gamma}_{i+1}}_{\mathsf{val}}(\text{Reach}(T))$;
2. if there is $i \geq 0$ such that $\gamma_i = \gamma_{i+1}$, then $\langle\!\langle 1\rangle\!\rangle^{\overline{\gamma}_i}_{\mathsf{val}}(\text{Reach}(T)) = \langle\!\langle 1\rangle\!\rangle_{\mathsf{val}}(\text{Reach}(T))$; and
3. $\lim_{i\to\infty} \langle\!\langle 1\rangle\!\rangle^{\overline{\gamma}_i}_{\mathsf{val}}(\text{Reach}(T)) = \langle\!\langle 1\rangle\!\rangle_{\mathsf{val}}(\text{Reach}(T))$.

Condition 1 guarantees that the algorithm computes a sequence of monotonically improving selectors. Condition 2 guarantees that if a selector cannot be improved, then it is optimal. Condition 3 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 $\overline{\gamma}_i$ is $\varepsilon$-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 $\varepsilon > 0$ [14, 10]. For *turn-based* reachability games, our algorithm terminates with an optimal selector and we will present bounds for termination.

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 $\eta_0, \eta_1, \eta_2, \ldots$ computed in Section 4.1 may violate Condition 2: 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

**Algorithm 1** Reachability Strategy-Improvement Algorithm

**Input:** a concurrent game structure $G$ with target set $T$.
**Output:** a strategy $\overline{\gamma}$ for player 1.

0. Compute $W_2 = \{s \in S \mid \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\text{Reach}(T))(s) = 0\}$.
1. Let $\gamma_0 = \xi_1^{unif}$ and $i = 0$.
2. Compute $v_0 = \langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_0}(\text{Reach}(T))$.
3. **do** $\{$
   3.1. Let $I = \{s \in S \setminus (T \cup W_2) \mid Pre_1(v_i)(s) > v_i(s)\}$.
   3.2. Let $\xi_1$ be a player-1 selector such that for all states $s \in I$, we have $Pre_{1:\xi_1}(v_i)(s) = Pre_1(v_i)(s) > v_i(s)$.
   3.3. The player-1 selector $\gamma_{i+1}$ is defined as follows: for each state $s \in S$, let
$$\gamma_{i+1}(s) = \begin{cases} \gamma_i(s) & \text{if } s \notin I; \\ \xi_1(s) & \text{if } s \in I. \end{cases}$$
   3.4. Compute $v_{i+1} = \langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_{i+1}}(\text{Reach}(T))$.
   3.5. Let $i = i + 1$.

$\}$ **until** $I = \emptyset$.

4. **return** $\overline{\gamma}_i$.

---

follows: given two player 1 strategies $\pi_1$ and $\pi_1'$, let $\pi_1 \prec \pi_1'$ if the following two conditions hold: (i) $\langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\pi_1}(\text{Reach}(T)) \leq \langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\pi_1'}(\text{Reach}(T))$; and (ii) $\langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\pi_1}(\text{Reach}(T))(s) < \langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\pi_1'}(\text{Reach}(T))(s)$ for some state $s \in S$. Furthermore, we write $\pi_1 \preceq \pi_1'$ 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 $\gamma_0 = \overline{\xi}_1^{unif}$. At iteration $i+1$, the algorithm considers the memoryless player-1 strategy $\overline{\gamma}_i$ and computes the value $\langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_i}(\text{Reach}(T))$. Observe that since $\overline{\gamma}_i$ is a memoryless strategy, the computation of $\langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_i}(\text{Reach}(T))$ involves solving the 2-MDP $G_{\gamma_i}$. The valuation $\langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_i}(\text{Reach}(T))$ is named $v_i$. For all states $s$ such that $Pre_1(v_i)(s) > v_i(s)$, 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 $Pre_1(v_i) = v_i$, the algorithm stops and returns the optimal memoryless strategy $\overline{\gamma}_i$ for player 1. Unlike strategy-improvement algorithms for turn-based games (see [6] 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 $U_r^i = \{s \in S \setminus (T \cup W_2) \mid \langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_i}(\text{Reach}(T))(s) = v_i(s) = r\}$ be the value class with value $r$ at iteration $i$. For a state $s \in U_r^i$ the following assertion holds: if $Dest(s, \gamma_i, \xi_2) \subsetneq U_r^i$, then $Dest(s, \gamma_i, \xi_2) \cap U_{>r}^i \neq \emptyset$. Let $z = \max\{r \mid U_r^i \cap C \neq \emptyset\}$, that is, $U_z^i$ 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 $s \in U_q^i$. Since $Pre_1(v_i)(s) > v_i(s)$, it follows that $Dest(s, \gamma_{i+1}, \xi_2) \cap U_{>q}^i \neq \emptyset$. Hence we must have $z > q$, and therefore $I \cap C \cap U_z^i = \emptyset$. Thus, for all states $s \in U_z^i \cap C$, 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 $U_{>z}^i \cap C = \emptyset$. Thus for all states $s \in C \cap U_z^i$, we have $Dest(s, \gamma_{i+1}, \xi_2) \subseteq U_z^i \cap C$. It follows that $C \subseteq U_z^i$. 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), $I \cap C \cap U_z^i = \emptyset$, and $C \subseteq U_z^i$. ∎

**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 $I = \{s \in S \mid Pre_1(v_i)(s) > v_i(s)\}$. Let $v_i = \langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_i}(\textit{Reach}(T))$ and $v_{i+1} = \langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_{i+1}}(\textit{Reach}(T))$. Then $v_{i+1}(s) \geq Pre_1(v_i)(s)$ for all states $s \in S$; and therefore $v_{i+1}(s) \geq v_i(s)$ for all states $s \in S$, and $v_{i+1}(s) > v_i(s)$ for all states $s \in I$.*

**Proof.** 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

$$\widehat{w}_i(s) = \begin{cases} w_i(s) & \text{if } s \in S \setminus I; \\ 1 - Pre_1(v_i)(s) < w_i(s) & \text{if } s \in I. \end{cases}$$

In other words, $\widehat{w}_i = 1 - Pre_1(v_i)$, and we also have $\widehat{w}_i \leq w_i$. We now show that $\widehat{w}_i$ is a feasible solution to the linear program for MDPs with the objective $\text{Reach}(W_2)$, as described in Section 3. Since $v_i = \langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_i}(\text{Reach}(T))$, it follows that for all states $s \in S$ and all moves $a_2 \in \Gamma_2(s)$, we have

$$w_i(s) \geq \sum_{t \in S} w_i(t) \cdot \delta_{\gamma_i}(s, a_2).$$

For all states $s \in S \setminus I$, we have $\gamma_i(s) = \gamma_{i+1}(s)$ and $\widehat{w}_i(s) = w_i(s)$, and since $\widehat{w}_i \leq w_i$, it follows that for all states $s \in S \setminus I$ and all moves $a_2 \in \Gamma_2(s)$, we have

$$\widehat{w}_i(s) \geq \sum_{t \in S} \widehat{w}_i(t) \cdot \delta_{\gamma_{i+1}}(s, a_2) \qquad (\text{for } s \in (S \setminus I)).$$

Since for $s \in I$ the selector $\gamma_{i+1}(s)$ is obtained as an optimal selector for $Pre_1(v_i)(s)$, it follows that for all states $s \in I$ and all moves $a_2 \in \Gamma_2(s)$, we have

$$Pre_{\gamma_{i+1},a_2}(v_i)(s) \geq Pre_1(v_i)(s);$$

in other words, $1 - Pre_1(v_i)(s) \geq 1 - Pre_{\gamma_{i+1},a_2}(v_i)(s)$. Hence for all states $s \in I$ and all moves $a_2 \in \Gamma_2(s)$, we have

$$\widehat{w}_i(s) \geq \sum_{t \in S} w_i(t) \cdot \delta_{\gamma_{i+1}}(s, a_2).$$

Since $\widehat{w}_i \leq w_i$, for all states $s \in I$ and all moves $a_2 \in \Gamma_2(s)$, we have

$$\widehat{w}_i(s) \geq \sum_{t \in S} \widehat{w}_i(t) \cdot \delta_{\gamma_{i+1}}(s, a_2) \qquad (\text{ for } s \in I).$$

Hence it follows that $\widehat{w}_i$ 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 $v_{i+1} \geq 1 - \widehat{w}_i = Pre_1(v_i)$. 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 $s \in I$. ∎

**Theorem 3** (STRATEGY IMPROVEMENT). *The following two assertions hold about Algorithm 1:*

1. *For all $i \geq 0$, we have $\overline{\gamma}_i \preceq \overline{\gamma}_{i+1}$; moreover, if $\overline{\gamma}_i = \overline{\gamma}_{i+1}$, then $\overline{\gamma}_i$ is an optimal strategy.*
2. $\lim_{i\to\infty} v_i = \lim_{i\to\infty} \langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_i}(\textit{Reach}(T)) = \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\textit{Reach}(T))$.

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

1. The assertion that $\overline{\gamma}_i \preceq \overline{\gamma}_{i+1}$ follows from Lemma 8. If $\overline{\gamma}_i = \overline{\gamma}_{i+1}$, then $Pre_1(v_i) = v_i$. Let $v = \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\text{Reach}(T))$, and since $v$ is the least solution to satisfy $Pre_1(x) = x$ (i.e., the least fixpoint) [11], it follows that $v_i \geq v$. From Lemma 7 it follows that $\overline{\gamma}_i$ is proper. Since $\overline{\gamma}_i$ is proper by Lemma 3, we have $\langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_i}(\text{Reach}(T)) \geq v_i \geq v$. It follows that $\overline{\gamma}_i$ 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 $v_{k+1} \geq [T] \vee Pre_1(v_k)$. For all $k \geq 0$, let $u_{k+1} = [T] \vee Pre_1(u_k)$. By induction we conclude that for all $k \geq 0$, we have $u_k \leq v_k$. Moreover, $v_k \leq \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\text{Reach}(T))$, that is, for all $k \geq 0$, we have

$$u_k \leq v_k \leq \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\text{Reach}(T)).$$

   Since $\lim_{k\to\infty} u_k = \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\text{Reach}(T))$, it follows that

$$\lim_{k\to\infty} \langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_k}(\text{Reach}(T)) = \lim_{k\to\infty} v_k = \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\text{Reach}(T)).$$

The theorem follows. ∎

### 5.3 Termination for turn-based stochastic games

If the input game structure to Algorithm 1 is a turn-based stochastic game structure, then if we start with a proper selector $\gamma_0$ that is pure, then for all $i \geq 0$ we can choose the selector $\gamma_i$ such that $\gamma_i$ is both proper and pure: the above claim follows since given a valuation $v$, if a state $s$ is a player 1 state, then there is an action $a$ at $s$ (or choice of an edge at $s$) that achieves $Pre_1(v)(s)$ at $s$. Since the number of pure selectors is bounded, if we start with a pure, proper selector then termination is ensured. Hence we present a procedure to compute a pure, proper selector, and then present termination bounds (i.e., bounds on $i$ such that $u_{i+1} = u_i$). The construction of a pure, proper selector is based on the notion of *attractors* defined below.

*Attractor strategy.* Let $A_0 = W_2 \cup T$, and for $i \geq 0$ we have

$$A_{i+1} = A_i \cup \{s \in S_1 \cup S_R \mid E(s) \cap A_i \neq \emptyset\} \cup \{s \in S_2 \mid E(s) \subseteq A_i\}.$$

Since for all $s \in S \setminus W_2$ we have $\langle\langle 1 \rangle\rangle_{\mathsf{val}}(\text{Reach}(T)) > 0$, it follows that from all states in $S \setminus W_2$ player 1 can ensure that $T$ is reached with positive probability. It follows that for some $i \geq 0$ we have $A_i = S$. The pure *attractor* selector $\xi^*$ is as follows: for a state $s \in (A_{i+1} \setminus A_i) \cap S_1$ we have $\xi^*(s)(t) = 1$, where $t \in A_i$ (such a $t$ exists by construction). The pure memoryless strategy $\overline{\xi^*}$ ensures that for all $i \geq 0$, from $A_{i+1}$ the game reaches $A_i$ with positive probability. Hence there is no end-component $C$ contained in $S \setminus (W_2 \cup T)$ in the MDP $G_{\overline{\xi^*}}$. It follows that $\xi^*$ is a pure selector that is proper, and the selector $\xi^*$ can be computed in $O(|E|)$ time. We now present the termination bounds.

*Termination bounds.* We present termination bounds for binary turn-based stochastic games. A turn-based stochastic game is binary if for all $s \in S_R$ we have $|E(s)| \leq 2$, and for all $s \in S_R$ if $|E(s)| = 2$, then for all $t \in E(s)$ we have $\delta(s)(t) = \frac{1}{2}$, i.e., for all probabilistic states there are at most two successors and the transition function $\delta$ is uniform.

**Lemma 9** *Let $G$ be a binary Markov chain with $|S|$ states with a reachability objective Reach$(T)$. Then for all $s \in S$ we have $\langle\langle 1 \rangle\rangle_{\mathsf{val}}(\text{Reach}(T)) = \frac{p}{q}$, with $p, q \in \mathbb{N}$ and $p, q \leq 4^{|S|-1}$.*

**Proof.** The results follow as a special case of Lemma 2 of [6]. Lemma 2 of [6] holds for halting turn-based stochastic games, and since Markov chains reaches the set of closed connected recurrent states with probability 1 from all states the result follows. ∎

**Lemma 10** *Let $G$ be a binary turn-based stochastic game with a reachability objective Reach$(T)$. Then for all $s \in S$ we have $\langle\langle 1 \rangle\rangle_{\mathsf{val}}(\text{Reach}(T)) = \frac{p}{q}$, with $p, q \in \mathbb{N}$ and $p, q \leq 4^{|S_R|-1}$.*

**Proof.** Since pure memoryless optimal strategies exist for both players (existence of pure memoryless optimal strategies for both players in turn-based stochastic reachability games follows from [5]), we fix pure memoryless optimal strategies $\pi_1$ and $\pi_2$ for both players. The Markov chain $G_{\pi_1,\pi_2}$ can be then reduced to an equivalent Markov chains with $|S_R|$ states (since we fix deterministic successors for states in $S_1 \cup S_2$, they can be collapsed to their successors). The result then follows from Lemma 9. ∎

From Lemma 10 it follows that at iteration $i$ of the reachability strategy improvement algorithm either the sum of the values either increases by $\frac{1}{4^{|S_R|-1}}$ or else there is a valuation $u_i$ such that $u_{i+1} = u_i$. Since the sum of values of all states can be at most $|S|$, it follows that algorithm terminates in at most $|S| \cdot 4^{|S_R|-1}$ iterations. Moreover, since the number of pure memoryless strategies is at most $\prod_{s \in S_1} |E(s)|$, the algorithm terminates in at most $\prod_{s \in S_1} |E(s)|$ iterations. It follows from the results of [28] that a turn-based stochastic game structure $G$ can be reduced to a equivalent binary turn-based stochastic game structure $G'$ such that the set of player 1 and player 2 states in $G$ and $G'$ are the same and the number of probabilistic states in $G'$ is $O(|\delta|)$, where $|\delta|$ is the size of the transition function in $G$. Thus we obtain the following result.

**Theorem 4** *Let $G$ be a turn-based stochastic game with a reachability objective Reach$(T)$, then the reachability strategy improvement algorithm computes the values in time*

$$O\big(\min\{\prod_{s \in S_1} |E(s)|, 2^{O(|\delta|)}\} \cdot poly(|G|)\big);$$

*where $poly$ is polynomial function.*

The results of [16] presented an algorithm for turn-based stochastic games that works in time $O(|S_R|! \cdot poly(|G|))$. The algorithm of [16] works only for turn-based stochastic games, for general turn-based stochastic games the complexity of the algorithm of [16] is better. However, for turn-based stochastic games where the transition function at all states can be expressed with constantly many bits we have $|\delta| = O(|S_R|)$. In these cases the reachability strategy improvement algorithm (that works for both concurrent and turn-based stochastic games) works in time $2^{O(|S_R|)} \cdot poly(|G|)$ as compared to the time $2^{O(|S_R| \cdot \log(|S_R|)} \cdot poly(|G|)$ of the algorithm of [16].

# 6 Existence of Memoryless Optimal Strategies for Concurrent Safety Games

A proof of the existence of memoryless optimal strategies for safety games can be found in [11]: the proof uses results on martingales to obtain the result. For sake of completeness we present (an alternative) proof of the result: the proof we present is similar in spirit with the other proofs in this paper and uses the results on MDPs to obtain the result. The proof is very similar to the proof presented in [13].

**Theorem 5** (MEMORYLESS OPTIMAL STRATEGIES). *Memoryless optimal strategies exist for all concurrent games with safety objectives.*

**Proof.** Consider a concurrent game structure $G$ with an safety objective Safe($F$) for player 1. Then it follows from the results of [11] that

$$\langle\langle 1\rangle\rangle_{\mathsf{val}}(\text{Safe}(F)) = \nu X.\big(\min\{[F], Pre_1(X)\}\big),$$

where $[F]$ is the indicator function of the set $F$ and $\nu$ denotes the greatest fixpoint. Let $T = S \setminus F$, and for all states $s \in T$ we have $\langle\langle 1\rangle\rangle_{\mathsf{val}}(\text{Safe}(F))(s) = 0$, and hence any memoryless strategy from $T$ is an optimal strategy. Thus without loss of generality we assume all states in $T$ are absorbing. Let $v = \langle\langle 1\rangle\rangle_{\mathsf{val}}(\text{Safe}(F))$, and since we assume all states in $T$ are absorbing it follows that $Pre_1(v) = v$ (since $v$ is a fixpoint). Let $\gamma$ be a player 1 selector such that for all states $s$ we have $Pre_{1:\gamma}(v)(s) = Pre_1(v)(s) = v(s)$. We show that $\overline{\gamma}$ is an memoryless optimal strategy. Consider the player-2 MDP $G_\gamma$ and we consider the maximal probability for player 2 to reach the target set $T$. Consider the valuation $w$ defined as $w = 1 - v$. For all states $s \in T$ we have $w(s) = 1$. Since $Pre_{1:\gamma}(v) = Pre_1(v)$ it follows that for all states $s \in F$ and all $a_2 \in \Gamma_2(s)$ we have

$$Pre_{\gamma,a_2}(v)(s) \geq Pre_1(v)(s) = v(s);$$

in other words, for all $s \in F$ we have $1 - Pre_1(v)(s) = 1 - v(s) \geq 1 - Pre_{\gamma,a_2}(v)(s)$. Hence for all states $s \in F$ and all moves $a_2 \in \Gamma_2(s)$, we have

$$w(s) \geq \sum_{t \in S} w(t) \cdot \delta_\gamma(s, a_2).$$

Hence it follows that $w$ is a feasible solution to the linear program for MDPs with reachability objectives, i.e., given the memoryless strategy $\overline{\gamma}$ for player 1 the maximal probability valuation for player 2 to reach $T$ is at most $w$. Hence the memoryless strategy $\overline{\gamma}$ ensures that the probability valaution for player 1 to stay safe in $F$ against all player 2 strategies is at least $v = \langle\langle 1\rangle\rangle_{\mathsf{val}}(\text{Safe}(F))$. Optimality of $\overline{\gamma}$ follows. ∎

# 7 Strategy Improvement Algorithm for Concurrent Safety Games

In this section we present a strategy improvement algorithm for concurrent games with safety objectives. We consider a concurrent game structure with a safe set $F$, i.e., the objective for player 1 is Safe$(F)$. The algorithm will produce a sequence of selectors $\gamma_0, \gamma_1, \gamma_2, \ldots$ for player 1, such that Condition 1, Condition 2 and Condition 3 of Section 5 are satisfied. Note that for concurrent safety 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, as the value can be irrational [11]. We start with a few notations

**Optimal selectors.** Given a valuation $v$ and a state $s$, we define by

$$\mathsf{OptSel}(v, s) = \{\xi_1 \in \Lambda_1(s) \mid Pre_{1:\xi_1}(v)(s) = Pre_1(v)(s)\}$$

the set of optimal selectors for $v$ at state $s$. For an optimal selector $\xi_1 \in \mathsf{OptSel}(v, s)$, we define the set of counter-optimal actions as follows:

$$\mathsf{CountOpt}(v, s, \xi_1) = \{b \in \Gamma_2(s) \mid Pre_{\xi_1,b}(v)(s) = Pre_1(v)(s)\}.$$

Observe that for $\xi_1 \in \mathsf{OptSel}(v, s)$, for all $b \in \Gamma_2(s) \setminus \mathsf{CountOpt}(v, s, \xi_1)$ we have $Pre_{\xi_1,b}(v)(s) > Pre_1(v)(s)$. We define the set of optimal selector support and the counter-optimal action set as follows:

$$\begin{array}{rcl} \mathsf{OptSelCount}(v, s) & = & \{(A, B) \subseteq \Gamma_1(s) \times \Gamma_2(s) \mid \exists \xi_1 \in \Lambda_1(s).\ \xi_1 \in \mathsf{OptSel}(v, s) \\ & & \wedge\ Supp(\xi_1) = A\ \wedge\ \mathsf{CountOpt}(v, s, \xi_1) = B\}; \end{array}$$

i.e., it consists of pairs $(A, B)$ of actions of player 1 and player 2, such that there is an optimal selector $\xi_1$ with support $A$, and $B$ is the set of counter-optimal actions to $\xi_1$.

**Turn-based reduction.** Given a concurrent game $G = \langle S, M, \Gamma_1, \Gamma_2, \delta\rangle$ and a valuation $v$ we construct a turn-based stochastic game $\overline{G}_v = \langle(\overline{S}, \overline{E}), (\overline{S}_1, \overline{S}_2, \overline{S}_R), \overline{\delta}\rangle$ as follows:

1. The set of states is as follows:

$$\begin{array}{rcl} \overline{S} & = & S \cup \{(s, A, B) \mid s \in S,\ (A, B) \in \mathsf{OptSelCount}(v, s)\} \\ & \cup & \{(s, A, b) \mid s \in S,\ (A, B) \in \mathsf{OptSelCount}(v, s),\ b \in B\}. \end{array}$$

2. The state space partition is as follows: $\overline{S}_1 = S$; $\overline{S}_2 = \{(s, A, B) \mid s \in S, (A, B) \in \mathsf{OptSelCount}(v, s)\}$; and $\overline{S}_R = \{(s, A, b) \mid s \in S,\ (A, B) \in \mathsf{OptSelCount}(v, s), b \in B\}$. In other words, $(\overline{S}_1, \overline{S}_2, \overline{S}_R)$ is a partition of the state space, where $\overline{S}_1$ are player 1 states, $\overline{S}_2$ are player 2 states, and $\overline{S}_R$ are random or probabilistic states.

3. The set of edges is as follows:

$$\begin{array}{rcl} \overline{E} & = & \{(s, (s, A, B)) \mid s \in S, (A, B) \in \mathsf{OptSelCount}(v, s)\} \\ & \cup & \{((s, A, B), (s, A, b)) \mid b \in B\} \cup \{((s, A, b), t) \mid t \in \bigcup_{a \in A} Dest(s, a, b)\}. \end{array}$$

4. The transition function $\overline{\delta}$ for all states in $\overline{S}_R$ is uniform over its successors.

Intuitively, the reduction is as follows. Given the valuation $v$, state $s$ is a player 1 state where player 1 can select a pair $(A, B)$ (and move to state $(s, A, B)$) with $A \subseteq \Gamma_1(s)$ and $B \subseteq \Gamma_2(s)$ such that there is an optimal selector $\xi_1$ with support exactly $A$ and the set of counter-optimal actions to $\xi_1$ is the set $B$. From a player 2 state $(s, A, B)$, player 2 can choose any action $b$ from the set $B$, and move to state $(s, A, b)$. A state $(s, A, b)$ is a probabilistic state where all the states in $\bigcup_{a \in A} Dest(s, a, b)$ are chosen uniformly at random. Given a set $F \subseteq S$ we denote by $\overline{F} = F \cup \{(s, A, B) \in \overline{S} \mid s \in F\} \cup \{(s, A, b) \in \overline{S} \mid s \in F\}$. We refer to the above reduction as $\mathsf{TB}$, i.e., $(\overline{G}_v, \overline{F}) = \mathsf{TB}(G, v, F)$.

**Value-class of a valuation.** Given a valuation $v$ and a real $0 \leq r \leq 1$, the *value-class* $U_r(v)$ of value $r$ is the set of states with valuation $r$, i.e., $U_r(v) = \{s \in S \mid v(s) = r\}$

## 7.1 The strategy-improvement algorithm

**Ordering of strategies.** Let $G$ be a concurrent game and $F$ be the set of safe states. Let $T = S \setminus F$. Given a concurrent game structure $G$ with a safety objective Safe$(F)$, the set of *almost-sure winning* states is the set of states $s$ such that the value at $s$ is 1, i.e., $W_1 = \{s \in S \mid \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\text{Safe}(F)) = 1\}$ is the set of almost-sure winning states. An optimal strategy from $W_1$ is referred as an almost-sure winning strategy. The set $W_1$ and an almost-sure winning strategy can be computed in linear time by the algorithm given in [9]. We assume without loss of generality that all states in $W_1 \cup T$ are absorbing. We recall the preorder $\prec$ on the strategies for player 1 (as defined in Section 5.1) 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) $\langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\pi_1}(\text{Safe}(F)) \leq \langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\pi_1'}(\text{Safe}(F))$; and (ii) $\langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\pi_1}(\text{Safe}(F))(s) < \langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\pi_1'}(\text{Safe}(F))(s)$ for some state $s \in S$. Furthermore, we write $\pi_1 \preceq \pi_1'$ if either $\pi_1 \prec \pi_1'$ or $\pi_1 = \pi_1'$. We first present an example that shows the improvements based only on $Pre_1$ operators are not sufficient for safety games, even on turn-based games and then present our algorithm.

**Example 2** Consider the turn-based stochastic game shown in Fig 2, where the $\Box$ states are player 1 states, the $\Diamond$ states are player 2 states, and $\bigcirc$ states are random states with probabilities labeled on edges. The safety goal is to avoid the state $s_4$. Consider a memoryless strategy $\pi_1$ for player 1 that chooses the successor $s_0 \to s_2$, and the counter-strategy $\pi_2$ for player 2 chooses $s_1 \to s_0$. Given the strategies $\pi_1$ and $\pi_2$, the value at $s_0, s_1$ and $s_2$ is $1/3$, and since all successors of $s_0$ have value $1/3$, the value cannot be improved by $Pre_1$. However, note that if player 2 is restricted to choose only value optimal selectors for the value $1/3$, then player 1 can switch to the strategy $s_0 \to s_1$ and ensure that the game stays in the value class $1/3$ with probability 1. Hence switching to $s_0 \to s_1$ would force player 2 to select a counter-strategy that switches to the strategy $s_1 \to s_3$, and thus player 1 can get a value $2/3$. ∎

**Informal description of Algorithm 2.** We first present the basic strategy improvement algorithm (Algorithm 2) and will later present a convergent version (Algorithm 4) for computing the values for all states in $S \setminus W_1$. The algorithm (Algorithm 2) iteratively improves player-1 strategies according to the preorder $\prec$. The algorithm starts with the random selector $\gamma_0 = \overline{\xi}_1^{unif}$ that plays at all states all actions uniformly at random. At iteration $i+1$, the algorithm considers the memoryless player-1 strategy $\overline{\gamma_i}$ and computes the value $\langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\overline{\gamma_i}}(\text{Safe}(F))$. Observe that since $\overline{\gamma_i}$ is a memoryless strategy, the computation of $\langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\overline{\gamma_i}}(\text{Safe}(F))$ involves solving the 2-MDP $G_{\gamma_i}$. The valuation $\langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\overline{\gamma_i}}(\text{Safe}(F))$ is named $v_i$. For all states $s$ such that $Pre_1(v_i)(s) > v_i(s)$, 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 $Pre_1(v_i) = v_i$, then the algorithm constructs the game $(\overline{G}_{v_i}, \overline{F}) = \mathsf{TB}(G, v_i, F)$, and computes $\overline{A}_i$ as the set of almost-sure winning states in $\overline{G}_{v_i}$ for the objective Safe$(\overline{F})$. Let $U = (\overline{A}_i \cap S) \setminus W_1$. If $U$ is non-empty, then a selector $\gamma_{i+1}$ is obtained at $U$ from an pure

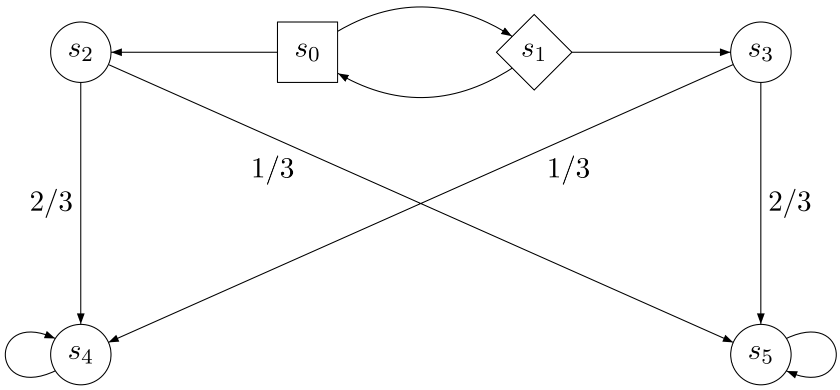


Figure 2: A turn-based stochastic safety game.

memoryless optimal strategy (i.e., an almost-sure winning strategy) in $\overline{G}_{v_i}$, and the algorithm proceeds to iteration $i+1$. If $Pre_1(v_i) = v_i$ and $U$ is empty, then the algorithm stops and returns the memoryless strategy $\overline{\gamma}_i$ for player 1. Unlike strategy improvement algorithms for turn-based games (see [6] for a survey), Algorithm 2 is not guaranteed to terminate (see Example 3). We will show that Algorithm 2 has both the monotonicity and optimality on termination properties, however, as we will illustrate in Example 3, the valuations of Algorithm 2 need not necessarily converge to the values. However, for turn-based stochastic games Algorithm 2 correctly converges to the values. We will show that Algorithm 4 has all the desired properties (i.e., monotonicity, optimality on termination, and convergence to the values).

**Lemma 11** *Let $\gamma_i$ and $\gamma_{i+1}$ be the player-1 selectors obtained at iterations $i$ and $i+1$ of Algorithm 2. Let $I = \{s \in S \setminus (W_1 \cup T) \mid Pre_1(v_i)(s) > v_i(s)\}$. Let $v_i = \langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_i}(Safe(F))$ and $v_{i+1} = \langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_{i+1}}(Safe(F))$. Then $v_{i+1}(s) \geq Pre_1(v_i)(s)$ for all states $s \in S$; and therefore $v_{i+1}(s) \geq v_i(s)$ for all states $s \in S$, and $v_{i+1}(s) > v_i(s)$ for all states $s \in I$.*

**Proof.** The proof is essentially similar to the proof of Lemma 8, and we present the details for completeness. 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 $s \in S$. The counter-optimal strategy for player 2 to minimize $v_{i+1}$ is obtained by maximizing the probability to reach $T$. Let

$$\widehat{w}_i(s) = \begin{cases} w_i(s) & \text{if } s \in S \setminus I; \\ 1 - Pre_1(v_i)(s) < w_i(s) & \text{if } s \in I. \end{cases}$$

In other words, $\widehat{w}_i = 1 - Pre_1(v_i)$, and we also have $\widehat{w}_i \leq w_i$. We now show that $\widehat{w}_i$ is a feasible solution to the linear program for MDPs with the objective $\text{Reach}(T)$, as described in Section 3. Since $v_i = \langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_i}(\text{Safe}(F))$, it follows that for all states $s \in S$ and all moves $a_2 \in \Gamma_2(s)$, we have

$$w_i(s) \geq \sum_{t \in S} w_i(t) \cdot \delta_{\gamma_i}(s, a_2).$$

For all states $s \in S \setminus I$, we have $\gamma_i(s) = \gamma_{i+1}(s)$ and $\widehat{w}_i(s) = w_i(s)$, and since $\widehat{w}_i \leq w_i$, it follows that for all states $s \in S \setminus I$ and all moves $a_2 \in \Gamma_2(s)$, we have

$$\widehat{w}_i(s) = w_i(s) \geq \sum_{t \in S} \widehat{w}_i(t) \cdot \delta_{\gamma_{i+1}}(s, a_2) \qquad (\text{ for } s \in S \setminus I).$$

**Algorithm 2** Safety Strategy-Improvement Algorithm

**Input:** a concurrent game structure $G$ with safe set $F$.
**Output:** a strategy $\overline{\gamma}$ for player 1.
0. Compute $W_1 = \{s \in S \mid \langle\langle 1\rangle\rangle_{\mathsf{val}}(\mathsf{Safe}(F))(s) = 1\}$.
1. Let $\gamma_0 = \xi_1^{unif}$ and $i = 0$.
2. Compute $v_0 = \langle\langle 1\rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_0}(\mathsf{Safe}(F))$.
3. **do** $\{$
3.1. Let $I = \{s \in S \setminus (W_1 \cup T) \mid Pre_1(v_i)(s) > v_i(s)\}$.
3.2 **if** $I \neq \emptyset$, **then**
3.2.1 Let $\xi_1$ be a player-1 selector such that for all states $s \in I$, we have $Pre_{1:\xi_1}(v_i)(s) = Pre_1(v_i)(s) > v_i(s)$.
3.2.2 The player-1 selector $\gamma_{i+1}$ is defined as follows: for each state $s \in S$, let
$$\gamma_{i+1}(s) = \begin{cases} \gamma_i(s) & \text{if } s \not\in I; \\ \xi_1(s) & \text{if } s \in I. \end{cases}$$
3.3 **else**
3.3.1 let$(\overline{G}_{v_i}, \overline{F}) = \mathsf{TB}(G, v_i, F)$
3.3.2 let $\overline{A}_i$ be the set of almost-sure winning states in $\overline{G}_{v_i}$ for Safe($\overline{F}$) and $\overline{\pi}_1$ be a pure memoryless almost-sure winning strategy from the set $\overline{A}_i$.
3.3.3 **if** $((\overline{A}_i \cap S) \setminus W_1 \neq \emptyset)$
3.3.3.1 let $U = (\overline{A}_i \cap S) \setminus W_1$
3.3.3.2 The player-1 selector $\gamma_{i+1}$ is defined as follows: for $s \in S$, let
$$\gamma_{i+1}(s) = \begin{cases} \gamma_i(s) & \text{if } s \not\in U; \\ \xi_1(s) & \text{if } s \in U, \xi_1(s) \in \mathsf{OptSel}(v_i, s), \\ & \overline{\pi}_1(s) = (s, A, B), B = \mathsf{OptSelCount}(s, v, \xi_1). \end{cases}$$
3.4. Compute $v_{i+1} = \langle\langle 1\rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_{i+1}}(\mathsf{Safe}(F))$.
3.5. Let $i = i + 1$.
$\}$ **until** $I = \emptyset$ and $(\overline{A}_{i-1} \cap S) \setminus W_1 = \emptyset$.
4. **return** $\overline{\gamma}_i$.

Since for $s \in I$ the selector $\gamma_{i+1}(s)$ is obtained as an optimal selector for $Pre_1(v_i)(s)$, it follows that for all states $s \in I$ and all moves $a_2 \in \Gamma_2(s)$, we have
$$Pre_{\gamma_{i+1},a_2}(v_i)(s) \geq Pre_1(v_i)(s);$$
in other words, $1 - Pre_1(v_i)(s) \geq 1 - Pre_{\gamma_{i+1},a_2}(v_i)(s)$. Hence for all states $s \in I$ and all moves $a_2 \in \Gamma_2(s)$, we have
$$\widehat{w}_i(s) \geq \sum_{t \in S} w_i(t) \cdot \delta_{\gamma_{i+1}}(s, a_2).$$
Since $\widehat{w}_i \leq w_i$, for all states $s \in I$ and all moves $a_2 \in \Gamma_2(s)$, we have
$$\widehat{w}_i(s) \geq \sum_{t \in S} \widehat{w}_i(t) \cdot \delta_{\gamma_{i+1}}(s, a_2) \qquad (\text{ for } s \in I).$$

Hence it follows that $\widehat{w}_i$ is a feasible solution to the linear program for MDPs with reachability objectives. Since the reachability valuation for player 2 for Reach$(T)$ is the least solution (observe that the objective function of the linear program is a minimizing function), it follows that $v_{i+1} \geq 1 - \widehat{w}_i = Pre_1(v_i)$. 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 $s \in I$. ∎

Recall that by Example 2 it follows that improvement by only step 3.2 is not sufficient to guarantee convergence to optimal values. We now present a lemma about the turn-based reduction, and then show that step 3.3 also leads to an improvement. Finally, in Theorem 7 we show that if improvements by step 3.2 and step 3.3 are not possible, then the optimal value and an optimal strategy is obtained.

**Lemma 12** *Let $G$ be a concurrent game with a set $F$ of safe states. Let $v$ be a valuation and consider $(\overline{G}_v, \overline{F}) = \mathsf{TB}(G, v, F)$. Let $\overline{A}$ be the set of almost-sure winning states in $\overline{G}_v$ for the objective Safe$(\overline{F})$, and let $\overline{\pi}_1$ be a pure memoryless almost-sure winning strategy from $\overline{A}$ in $\overline{G}_v$. Consider a memoryless strategy $\pi_1$ in $G$ for states in $\overline{A} \cap S$ as follows: if $\overline{\pi}_1(s) = (s, A, B)$, then $\pi_1(s) \in \mathsf{OptSel}(v, s)$ such that $Supp(\pi_1(s)) = A$ and $\mathsf{OptSelCount}(v, s, \pi_1(s)) = B$. Consider a pure memoryless strategy $\pi_2$ for player 2. If for all states $s \in \overline{A} \cap S$, we have $\pi_2(s) \in \mathsf{OptSelCount}(v, s, \pi_1(s))$, then for all $s \in \overline{A} \cap S$, we have $\Pr_s^{\pi_1, \pi_2}($Safe$(F)) = 1$.*

**Proof.** We analyze the Markov chain arising after the player fixes the memoryless strategies $\pi_1$ and $\pi_2$. Given the strategy $\pi_2$ consider the strategy $\overline{\pi}_2$ as follows: if $\overline{\pi}_1(s) = (s, A, B)$ and $\pi_2(s) = b \in \mathsf{OptSelCount}(v, s, \pi_1(s))$, then at state $(s, A, B)$ choose the successor $(s, A, b)$. Since $\overline{\pi}_1$ is an almost-sure winning strategy for Safe$(\overline{F})$, it follows that in the Markov chain obtained by fixing $\overline{\pi}_1$ and $\overline{\pi}_2$ in $\overline{G}_v$, all closed connected recurrent set of states that intersect with $\overline{A}$ are contained in $\overline{A}$, and from all states of $\overline{A}$ the closed connected recurrent set of states within $\overline{A}$ are reached with probability 1. It follows that in the Markov chain obtained from fixing $\pi_1$ and $\pi_2$ in $G$ all closed connected recurrent set of states that intersect with $\overline{A} \cap S$ are contained in $\overline{A} \cap S$, and from all states of $\overline{A} \cap S$ the closed connected recurrent set of states within $\overline{A} \cap S$ are reached with probability 1. The desired result follows. ∎

**Lemma 13** *Let $\gamma_i$ and $\gamma_{i+1}$ be the player-1 selectors obtained at iterations $i$ and $i+1$ of Algorithm 2. Let $I = \{s \in S \setminus (W_1 \cup T) \mid Pre_1(v_i)(s) > v_i(s)\} = \emptyset$, and $(\overline{A}_i \cap S) \setminus W_1 \neq \emptyset$. Let $v_i = \langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_i}($Safe$(F))$ and $v_{i+1} = \langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_{i+1}}($Safe$(F))$. Then $v_{i+1}(s) \geq v_i(s)$ for all states $s \in S$, and $v_{i+1}(s) > v_i(s)$ for some state $s \in (\overline{A}_i \cap S) \setminus W_1$.*

**Proof.** We first show that $v_{i+1} \geq v_i$. Let $U = (\overline{A}_i \cap S) \setminus W_1$. Let $w_i(s) = 1 - v_i(s)$ for all states $s \in S$. Since $v_i = \langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_i}(\mathrm{Safe}(F))$, it follows that for all states $s \in S$ and all moves $a_2 \in \Gamma_2(s)$, we have

$$w_i(s) \geq \sum_{t \in S} w_i(t) \cdot \delta_{\gamma_i}(s, a_2).$$

The selector $\xi_1(s)$ chosen for $\gamma_{i+1}$ at $s \in U$ satisfies that $\xi_1(s) \in \mathsf{OptSel}(v_i, s)$. It follows that for all states $s \in S$ and all moves $a_2 \in \Gamma_2(s)$, we have

$$w_i(s) \geq \sum_{t \in S} w_i(t) \cdot \delta_{\gamma_{i+1}}(s, a_2).$$

It follows that the maximal probability with which player 2 can reach $T$ against the strategy $\overline{\gamma}_{i+1}$ is at most $w_i$. It follows that $v_i(s) \leq v_{i+1}(s)$.

We now argue that for some state $s \in U$ we have $v_{i+1}(s) > v_i(s)$. Given the strategy $\overline{\gamma}_{i+1}$, consider a pure memoryless counter-optimal strategy $\pi_2$ for player 2 to reach $T$. Since the selectors $\gamma_{i+1}(s)$ at states $s \in U$ are obtained from the almost-sure strategy $\overline{\pi}$ in the turn-based game $\overline{G}_{v_i}$ to satisfy $\mathsf{Safe}(\overline{F})$, it follows from Lemma 12 that if for every state $s \in U$, the action $\pi_2(s) \in \mathsf{OptSelCount}(v_i, s, \gamma_{i+1})$, then from all states $s \in U$, the game stays safe in $F$ with probability 1. Since $\overline{\gamma}_{i+1}$ is a given strategy for player 1, and $\pi_2$ is counter-optimal against $\overline{\gamma}_{i+1}$, this would imply that $U \subseteq \{s \in S \mid \langle\langle 1\rangle\rangle_{\mathsf{val}}(\mathsf{Safe}(F)) = 1\}$. This would contradict that $W_1 = \{s \in S \mid \langle\langle 1\rangle\rangle_{\mathsf{val}}(\mathsf{Safe}(F)) = 1\}$ and $U \cap W_1 = \emptyset$. It follows that for some state $s^* \in U$ we have $\pi_2(s^*) \notin \mathsf{OptSelCount}(v_i, s^*, \gamma_{i+1})$, and since $\gamma_{i+1}(s^*) \in \mathsf{OptSel}(v_i, s^*)$ we have

$$v_i(s^*) < \sum_{t \in S} v_i(t) \cdot \delta_{\gamma_{i+1}}(s^*, \pi_2(s^*));$$

in other words, we have

$$w_i(s^*) > \sum_{t \in S} w_i(t) \cdot \delta_{\gamma_{i+1}}(s^*, \pi_2(s^*)).$$

Define a valuation $z$ as follows: $z(s) = w_i(s)$ for $s \neq s^*$, and $z(s^*) = \sum_{t \in S} w_i(t) \cdot \delta_{\gamma_{i+1}}(s^*, \pi_2(s^*))$. Given the strategy $\overline{\gamma}_{i+1}$ and the counter-optimal strategy $\pi_2$, the valuation $z$ satisfies the inequalities of the linear-program for reachability to $T$. It follows that the probability to reach $T$ given $\overline{\gamma}_{i+1}$ is at most $z$. Thus we obtain that $v_{i+1}(s) \geq v_i(s)$ for all $s \in S$, and $v_{i+1}(s^*) > v_i(s^*)$. This concludes the proof. ∎

We obtain the following theorem from Lemma 11 and Lemma 13 that shows that the sequences of values we obtain is monotonically non-decreasing.

**Theorem 6** (MONOTONICITY OF VALUES). *For $i \geq 0$, let $\gamma_i$ and $\gamma_{i+1}$ be the player-1 selectors obtained at iterations $i$ and $i+1$ of Algorithm 2. If $\gamma_i \neq \gamma_{i+1}$, then (a) for all $s \in S$ we have $\langle\langle 1\rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_i}(\mathit{Safe}(F))(s) \leq \langle\langle 1\rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_{i+1}}(\mathit{Safe}(F))(s)$; and (b) for some $s^* \in S$ we have $\langle\langle 1\rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_i}(\mathit{Safe}(F))(s^*) < \langle\langle 1\rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_{i+1}}(\mathit{Safe}(F))(s^*)$.*

**Theorem 7** (OPTIMALITY ON TERMINATION). *Let $v_i$ be the valuation at iteration $i$ of Algorithm 2 such that $v_i = \langle\langle 1\rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_i}(\mathit{Safe}(F))$. If $I = \{s \in S \setminus (W_1 \cup T) \mid Pre_1(v_i)(s) > v_i(s)\} = \emptyset$, and $(\overline{A}_i \cap S) \setminus W_1 = \emptyset$, then $\overline{\gamma}_i$ is an optimal strategy and $v_i = \langle\langle 1\rangle\rangle_{\mathsf{val}}(\mathit{Safe}(F))$.*

**Proof.** We show that for all memoryless strategies $\pi_1$ for player 1 we have $\langle\langle 1\rangle\rangle_{\mathsf{val}}^{\pi_1}(\mathsf{Safe}(F)) \leq v_i$. Since memoryless optimal strategies exist for concurrent games with safety objectives (Theorem 5) the desired result follows.

Let $\overline{\pi}_2$ be a pure memoryless optimal strategy for player 2 in $\overline{G}_{v_i}$ for the objective complementary to $\mathsf{Safe}(\overline{F})$, where $(\overline{G}_{v_i}, \mathsf{Safe}(\overline{F})) = \mathsf{TB}(G, v_i, F)$. Consider a memoryless strategy $\pi_1$ for player 1, and we define a pure memoryless strategy $\pi_2$ for player 2 as follows.

1. If $\pi_1(s) \notin \mathsf{OptSel}(v_i, s)$, then $\pi_2(s) = b \in \Gamma_2(s)$, such that $Pre_{\pi_1(s),b}(v_i)(s) < v_i(s)$; (such a $b$ exists since $\pi_1(s) \notin \mathsf{OptSel}(v_i, s)$).
2. If $\pi_1(s) \in \mathsf{OptSel}(v_i, s)$, then let $A = Supp(\pi_1(s))$, and consider $B$ such that $B = \mathsf{OptSelCount}(v_i, s, \pi_1(s))$. Then we have $\pi_2(s) = b$, such that $\overline{\pi}_2((s, A, B)) = (s, A, b)$.

Observe that by construction of $\pi_2$, for all $s \in S \setminus (W_1 \cup T)$, we have $Pre_{\pi_1(s),\pi_2(s)}(v_i)(s) \leq v_i(s)$. We first show that in the Markov chain obtained by fixing $\pi_1$ and $\pi_2$ in $G$, there is no closed connected recurrent set of states $C$ such that $C \subseteq S \setminus (W_1 \cup T)$. Assume towards contradiction that $C$ is a closed connected recurrent set of states in $S \setminus (W_1 \cup T)$. The following case analysis achieves the contradiction.

1. Suppose for every state $s \in C$ we have $\pi_1(s) \in \mathsf{OptSel}(v_i, s)$. Then consider the strategy $\overline{\pi}_1$ in $\overline{G}_{v_i}$ such that for a state $s \in C$ we have $\overline{\pi}_1(s) = (s, A, B)$, where $\pi_1(s) = A$, and $B = \mathsf{OptSelCount}(v_i, s, \pi_1(s))$. Since $C$ is closed connected recurrent states, it follows by construction that for all states $s \in C$ in the game $\overline{G}_{v_i}$ we have $\Pr_s^{\overline{\pi}_1, \overline{\pi}_2}(\mathsf{Safe}(\overline{C})) = 1$, where $\overline{C} = C \cup \{(s, A, B) \mid s \in C\} \cup \{(s, A, b) \mid s \in C\}$. It follows that for all $s \in C$ in $\overline{G}_{v_i}$ we have $\Pr_s^{\overline{\pi}_1, \overline{\pi}_2}(\mathsf{Safe}(\overline{F})) = 1$. Since $\overline{\pi}_2$ is an optimal strategy, it follows that $C \subseteq (\overline{A}_i \cap S) \setminus W_1$. This contradicts that $(\overline{A}_i \cap S) \setminus W_1 = \emptyset$.

2. Otherwise for some state $s^* \in C$ we have $\pi_1(s^*) \not\in \mathsf{OptSel}(v_i, s^*)$. Let $r = \min\{q \mid U_q(v_i) \cap C \neq \emptyset\}$, i.e., $r$ is the least value-class with non-empty intersection with $C$. Hence it follows that for all $q < r$, we have $U_q(v_i) \cap C = \emptyset$. Observe that since for all $s \in C$ we have $Pre_{\pi_1(s), \pi_2(s)}(v_i)(s) \leq v_i(s)$, it follows that for all $s \in U_r(v_i)$ either (a) $Dest(s, \pi_1(s), \pi_2(s)) \subseteq U_r(v_i)$; or (b) $Dest(s, \pi_1(s), \pi_2(s)) \cap U_q(v_i) \neq \emptyset$, for some $q < r$. Since $U_r(v_i)$ is the least value-class with non-empty intersection with $C$, it follows that for all $s \in U_r(v_i)$ we have $Dest(s, \pi_1(s), \pi_2(s)) \subseteq U_r(v_i)$. It follows that $C \subseteq U_r(v_i)$. Consider the state $s^* \in C$ such that $\pi_1(s^*) \not\in \mathsf{OptSel}(v_i, s)$. By the construction of $\pi_2(s)$, we have $Pre_{\pi_1(s^*), \pi_2(s^*)}(v_i)(s^*) < v_i(s^*)$. Hence we must have $Dest(s^*, \pi_1(s^*), \pi_2(s^*)) \cap U_q(v_i) \neq \emptyset$, for some $q < r$. Thus we have a contradiction.

It follows from above that there is no closed connected recurrent set of states in $S \setminus (W_1 \cup T)$, and hence with probability 1 the game reaches $W_1 \cup T$ from all states in $S \setminus (W_1 \cup T)$. Hence the probability to satisfy $\mathsf{Safe}(F)$ is equal to the probability to reach $W_1$. Since for all states $s \in S \setminus (W_1 \cup T)$ we have $Pre_{\pi_1(s), \pi_2(s)}(v_i)(s) \leq v_i(s)$, it follows that given the strategies $\pi_1$ and $\pi_2$, the valuation $v_i$ satisfies all the inequalities for linear program to reach $W_1$. It follows that the probability to reach $W_1$ from $s$ is atmost $v_i(s)$. It follows that for all $s \in S \setminus (W_1 \cup T)$ we have $\langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\pi_1}(\mathsf{Safe}(F))(s) \leq v_i(s)$. The result follows. ∎

**$k$-uniform selectors and strategies.** For concurrent games, we will use the result that for $\varepsilon > 0$, there is a $k$-*uniform memoryless* strategy that achieves the value of a safety objective within $\varepsilon$. We first define $k$-uniform selectors and $k$-uniform memoryless strategies. For a positive integer $k > 0$, a selector $\xi$ for player 1 is $k$-*uniform* if for all $s \in S \setminus (T \cup W_1)$ and all $a \in Supp(\pi_1(s))$ there exists $i, j \in \mathbb{N}$ such that $0 \leq i \leq j \leq k$ and $\xi(s)(a) = \frac{i}{j}$, i.e., the moves in the support are played with probability that are multiples of $\frac{1}{\ell}$ with $\ell \leq k$. We denote by $\Lambda^k$ the set of $k$-uniform selectors. A memoryless strategy is $k$-uniform if it is obtained from a $k$-uniform selector. We denote by $\Pi_1^{M,k}$ the set of $k$-uniform memoryless strategies for player 1. We first present a technical lemma (Lemma 14) that will be used in the key lemma (Lemma 15) to prove the convergence result.

**Lemma 14** *Let $a_1, a_2, \ldots, a_m$ be $m$ real numbers such that (1) for all $1 \leq i \leq m$, we have $a_i > 0$; and (2) $\sum_{i=1}^m a_i = 1$. Let $c = \min_{1 \leq i \leq m} a_i$. For $\eta > 0$, there exists $k \geq \frac{m}{c \cdot \eta}$ and $m$ real numbers $b_1, b_2, \ldots, b_m$ such that (1) for all $1 \leq i \leq m$, we have $b_i$ is a multiple of $\frac{1}{k}$ and $b_i > 0$; (2) $\sum_{i=1}^m b_i = 1$; and (3) for all $1 \leq i \leq m$, we have $\frac{a_i}{b_i} \leq 1 + \eta$ and $\frac{b_i}{a_i} \leq 1 + \eta$.*

**Proof.** Let $\ell = \frac{m}{\eta \cdot c}$. For $1 \leq i \leq m$, define $\overline{b}_i$ such that $\overline{b}_i$ is a multiple of $\frac{1}{\ell}$ and $a_i \leq \overline{b}_i \leq a_i + \frac{1}{\ell}$ (basically define $\overline{b}_i$ as the least multiple of $\frac{1}{\ell}$ that is at least the value of $a_i$). For $1 \leq i \leq m$, let $b_i = \frac{\overline{b}_i}{\sum_{i=1}^m \overline{b}_i}$; i.e., $b_i$ is defined from $\overline{b}_i$ with normalization. Clearly, $\sum_{i=1}^m b_i = 1$, and for all $1 \leq i \leq m$, we have $b_i > 0$ and $b_i$ can be expressed as a multiple of $\frac{1}{k}$, for some $k \geq \frac{m}{\eta \cdot c}$. We have the following inequalities: for all $1 \leq i \leq m$, we have

$$b_i \leq a_i + \frac{1}{\ell}; \qquad b_i \geq \frac{a_i}{1 + \frac{m}{\ell}}.$$

The first inequality follows since $\overline{b}_i \le a_i + \frac{1}{\ell}$ and $\sum_{i=1}^m \overline{b}_i \ge \sum_{i=1}^m a_i = 1$. The second inequality follows since $\overline{b}_i \ge a_i$ and $\sum_{i=1}^m \overline{b}_i \le \sum_{i=1}^m (a_i + \frac{1}{\ell}) = \sum_{i=1}^m a_i + \frac{m}{\ell} = 1 + \frac{m}{\ell}$. Hence for all $1 \le i \le m$, we have

$$\frac{b_i}{a_i} \le 1 + \frac{1}{\ell \cdot a_i} \le 1 + \frac{1}{\ell \cdot c} \le 1 + \eta;$$

$$\frac{a_i}{b_i} \le 1 + \frac{m}{\ell} \le 1 + \eta \cdot c \le 1 + \eta.$$

The desired result follows. ∎

**Lemma 15** *For all concurrent game structures $G$, for all safety objectives Safe$(F)$, for $F \subseteq S$, for all $\varepsilon > 0$, there exist $k > 0$ and $k$-uniform selectors $\xi$ such that $\overline{\xi}$ is an $\varepsilon$-optimal strategy.*

**Proof.** Our proof uses a result of Solan [27] and the existence of memoryless optimal strategies for concurrent safety games (Theorem 5). We first present the result of Solan specialized for MDPs with reachability objectives.

*The result of [27].* Let $G = (S, M, \Gamma_2, \delta)$ and $G' = (S, M, \Gamma_2, \delta')$ be two player-2 MDPs defined on the same state space $S$, with the same move set $M$ and the same move assignment function $\Gamma_2$, but with two different transition functions $\delta$ and $\delta'$, respectively. Let

$$\rho(G, G') = \max_{s,t \in S, a_2 \in \Gamma_2(s)} \left\{ \frac{\delta(s, a_2)(t)}{\delta'(s, a_2)(t)}, \frac{\delta'(s, a_2)(t)}{\delta(s, a_2)(t)} \right\} - 1;$$

where by convention $x/0 = +\infty$ for $x > 0$, and $0/0 = 1$ (compare with equation (9) of [27]: $\rho(G, G')$ is obtained as a specialization of (9) of [27] for MDPs). Let $T \subseteq S$. For $s \in S$, let $v(s)$ and $v'(s)$ denote the value for player 2 for the reachability objective Reach$(T)$ from $s$ in $G$ and $G'$, respectively. Then from Theorem 6 of [27] (also see equation (10) of [27]) it follows that

$$-4 \cdot |S| \cdot \rho(G, G') \le v(s) - v'(s) \le \frac{4 \cdot |S| \cdot \rho(G, G')}{(1 - 2 \cdot |S| \cdot \rho(G, G'))^+}; \tag{2}$$

where $x^+ = \max\{x, 0\}$. We first explain how specialization of Theorem 6 of [27] yields (2). Theorem 6 of [27] was proved for value functions of discounted games with costs, even when the discount factor $\lambda = 0$. Since the value functions of limit-average games are obtained as the limit of the value functions of discounted games as the discount factor goes to 0 [23], the result of Theorem 6 of [27] also holds for concurrent limit-average games (this was the main result of [27]). Since reachability objectives are special case of limit-average objectives, Theorem 6 of [27] also holds for reachability objectives. In the special case of reachability objectives with the same target set, the different cost functions used in equation (10) of [27] coincide, and the maximum absolute value of the cost is 1. Thus we obtain (2) as a specialization of Theorem 6 of [27].

We now use the existence of memoryless optimal strategies in concurrent safety games, and (2) to obtain our desired result. Consider a concurrent safety game $G = (S, M, \Gamma_1, \Gamma_2, \delta)$ with safe set $F$ for player 1. Let $\pi_1$ be a memoryless optimal strategy for the objective Safe$(F)$. Let $c = \min_{s \in S, a_1 \in \Gamma_1(s)} \{\pi_1(s)(a_1) \mid \pi_1(s)(a_1) > 0\}$ be the minimum positive transition probability given by $\pi_1$. Given $\varepsilon > 0$, let $\eta = \min\{\frac{1}{4 \cdot |S|}, \frac{\varepsilon}{8 \cdot |S|}\}$. We define a memoryless strategy $\pi_1'$ satisfying the following conditions: for $s \in S$ and $a_1 \in \Gamma_1(s)$ we have

1. if $\pi_1(s)(a_1) = 0$, then $\pi_1'(s)(a_1) = 0$;
2. if $\pi_1(s)(a_1) > 0$, then following conditions are satisfied:
   (a) $\pi_1'(s)(a_1) > 0$;
   (b) $\frac{\pi_1(s)(a_1)}{\pi_1'(s)(a_1)} \leq 1 + \eta$;
   (c) $\frac{\pi_1'(s)(a_1)}{\pi_1(s)(a_1)} \leq 1 + \eta$; and
   (d) $\pi_1'(s)(a_1)$ is a multiple of $\frac{1}{k}$, for an integer $k > 0$ (such a $k$ exists for $k > \frac{|M|}{c \cdot \eta}$).

For $k > \frac{|M|}{c \cdot \eta}$, such a strategy $\pi_1'$ exists (follows from the construction of Lemma 14). Let $G_1$ and $G_1'$ be the two player-2 MDPs obtained from $G$ by fixing the memoryless strategies $\pi_1$ and $\pi_1'$, respectively. Then by definition of $\pi_1'$ we have $\rho(G_1, G_1') \leq \eta$. Let $T = S \setminus F$. For $s \in S$, let the value of player 2 for the objective Reach$(T)$ in $G_1$ and $G_1'$ be $v(s)$ and $v'(s)$, respectively. By (2) we have

$$-4 \cdot |S| \cdot \eta \leq v(s) - v'(s) \leq \frac{4 \cdot |S| \cdot \eta}{(1 - 2 \cdot |S| \cdot \eta)^+};$$

Observe that by choice of $\eta$ we have (a) $4 \cdot |S| \cdot \eta \leq \frac{\varepsilon}{2 \cdot |S|}$ and (b) $2 \cdot |S| \cdot \eta \leq \frac{1}{2}$. Hence we have $-\varepsilon \leq v(s) - v'(s) \leq \varepsilon$. Since $\pi_1$ is a memoryless optimal strategy, it follows that $\pi_1'$ is a $k$-uniform memoryless $\varepsilon$-optimal strategy. ∎

**Turn-based stochastic games convergence.** We first observe that since pure memoryless optimal strategies exist for turn-based stochastic games with safety objectives (the results follows from results of [5, 22]), for turn-based stochastic games it suffices to iterate over pure memoryless selectors. Since the number of pure memoryless strategies is finite, it follows for turn-based stochastic games Algorithm 2 always terminates and yields an optimal strategy. In other words, we can restrict the selectors used in Algorithm 2 in Steps 3.2.2 and 3.3.2.2 to be pure memoryless selectors. Then the local improvement steps of Algorithm 2 with pure memoryless selectors terminates, and by Theorem 7 yield a globally optimal pure memoryless strategy (which is an optimal strategy). We will use the argument for turn-based stochastic games to a variant of Algorithm 2 restricted to $k$-uniform selectors.

**Strategy improvement with $k$-uniform selectors.** We now present the variant of Algorithm 2 where we restrict the algorithm to $k$-uniform selectors. The notations are essentially the same as used in Algorithm 2, but restricted to $k$-uniform selectors and presented as Algorithm 3. (for example, $\overline{G}_{v_i}^k$ is similar to $\overline{G}_{v_i}$ but restricted to $k$-uniform selectors, and similarly $\mathsf{OptSel}(v_i, s, k)$ are the optimal $k$-uniform selectors, see Section 8 for complete details). We first argue that if we restrict Algorithm 2 such that every iteration yields a $k$-uniform selector, for $k > 0$, then the algorithm terminates, i.e., Algorithm 3 terminates. The basic argument that if Algorithm 2 is restricted to $k$-uniform selectors for player 1, for $k > 0$, then the algorithm terminates, follows from the facts that (i) the sequence of strategies obtained are monotonic (Theorem 6) (i.e., the algorithm does not cycle among $k$-uniform selectors); and (ii) the number of $k$-uniform selectors for a given $k$ is finite. Given $k > 0$, let us denote by $z_i^k$ the valuation of Algorithm 3 at iteration $i$.

**Lemma 16** *For all $k > 0$, there exists $i \geq 0$ such that $z_i^k = z_{i+1}^k$.*

**Convergence to optimal $k$-uniform memoryless strategies.** We now argue that the valuation Algorithm 3 converges to is optimal for $k$-uniform selectors. The argument is as follows: if we restrict player 1 to chose between $k$-uniform selectors, then a concurrent game structures $G$ can be converted to a turn-based stochastic game structure, where player 1 first chooses a $k$-uniform selector, then player 2 chooses an action, and then the transition is determined by the chosen $k$-uniform selector of player 1, the action of player 2 and the transition function $\delta$ of the game structure $G$. Then by termination of turn-based stochastic games it follows that the algorithm will terminate. It follows from Theorem 7 that upon termination we obtain optimal strategy for the turn-based stochastic game. In other words, as discussed above for turn-based stochastic game, the local iteration converges to a globally optimal strategy. Hence the valuation obtained upon termination is the maximal value obtained over all $k$-uniform memoryless strategies. This gives us the following lemma (also see appendix for a detailed proof).

**Lemma 17** *For all $k > 0$, let $i \geq 0$ be such that $z_i^k = z_{i+1}^k$. Then we have $z_i^k = \max_{\pi_1 \in \Pi_1^{M,k}} \inf_{\pi_2 \in \Pi_2} \Pr^{\pi_1,\pi_2}(\mathit{Safe}(F))$.*

**Lemma 18** *For all concurrent game structures $G$, for all safety objectives $\mathit{Safe}(F)$, for $F \subseteq S$, for all $\varepsilon > 0$, there exist $k > 0$ and $i \geq 0$ such that for all $s \in S$ we have $z_i^k(s) \geq \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\mathit{Safe}(F))(s) - \varepsilon$.*

**Proof.** By Lemma 15, for all $\varepsilon > 0$, there exists $k > 0$ such that there is a $k$-uniform memoryless $\varepsilon$-optimal strategy for player 1. By Lemma 16, for all $k > 0$, there exists an $i \geq 0$ such that $z_i^k = z_{i+1}^k$, and by Lemma 17 it follows that the valuation $z_k^i$ represents the maximal value obtained by $k$-uniform memoryless strategies. Hence it follows that there exists $k > 0$ and $i \geq 0$ such that for all $s \in S$ we have $z_i^k(s) \geq \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\mathrm{Safe}(F))(s) - \varepsilon$. The desired result follows. ∎

We now present the convergent strategy improvement algorithm for safety objectives as Algorithm 4 that iterates over $k$-uniform strategy values. The algorithm iteratively calls Algorithm 3 with larger $k$, unless the termination condition of Algorithm 2 is satisfied.

**Theorem 8** (MONOTONICITY, OPTIMALITY ON TERMINATION AND CONVERGENCE). *Let $v_i$ be the valuation obtained at iteration $i$ of Algorithm 4. Then the following assertions hold.*

1. *For all $i \geq 0$ we have $v_{i+1} \geq v_i$.*
2. *If the algorithm terminates, then $v_i = \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\mathit{Safe}(F))$.*
3. *For all $\varepsilon > 0$, there exists $i$ such that for all $s$ we have $v_i(s) \geq \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\mathit{Safe}(F))(s) - \varepsilon$.*
4. $\lim_{i \to \infty} v_i = \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\mathit{Safe}(F))$.

**Proof.** We prove the results as follows.

1. Let $v_i$ is the valuation of Algorithm 4 at iteration $i$. For $k > 0$, we consider $z_i^k$ to denote the valuation of Algorithm 3 with the restriction of $k$-uniform selector at iteration $i$, and let $z_{i^*(k)}^k$ denote the least fixpoint (i.e., $i^*(k)$ is the least value of $i$ such that $z_i^k = z_{i+1}^k$). Since $k$-uniform selectors are a subset of $k+1$-uniform selectors, it follows that the maximal value obtained over strategies that uses $k+1$-uniform selectors is at least the maximal value obtained over $k$-uniform selectors. Since $z_{i^*(k)}^k$ denote the maximal value obtained over $k$-uniform selectors (follows from Lemma 17), we have that $z_{i^*(k)}^k \leq z_{i^*(k+1)}^{k+1}$ (note that we do not require that $i^*(k) \leq i^*(k+1)$, i.e., the algorithm with $k+1$-uniform selectors may require more iterations to terminate). We have $v_k = z_{i^*(k)}^k$ and hence the first result follows.

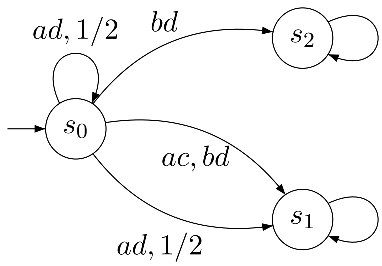


Figure 3: A simple game with irrational value.

2. The result follows from Theorem 7.

3. From Lemma 18 it follows that for all $\varepsilon > 0$, there exists a $k > 0$ such that for all $s$ we have $z^k_{i^*(k)}(s) \geq \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\mathrm{Safe}(F))(s) - \varepsilon$. Hence $v_k \geq \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\mathrm{Safe}(F))(s) - \varepsilon$. Hence we have that for all $\varepsilon > 0$, there exists $k \geq 0$, such that for all $s \in S$ we have $v_k(s) \geq \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\mathrm{Safe}(F))(s) - \varepsilon$.

4. By part (1) for all $i \geq 0$ we have $v_{i+1} \geq v_i$. By part (3), for all $\varepsilon > 0$, there exists $i \geq 0$ such that for all $s \in S$ we have $v_i(s) \geq \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\mathrm{Safe}(F))(s) - \varepsilon$. Hence it follows that for all $\varepsilon > 0$, there exists $i \geq 0$ such that for all $j \geq i$ and for all $s \in S$ we have $v_j(s) \geq \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\mathrm{Safe}(F))(s) - \varepsilon$. It follows that $\lim_{i \to \infty} v_i = \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\mathrm{Safe}(F))$.

This gives us the following result. ∎

**Discussion on convergence of Algorithm 2.** We will now present an example to illustrate that (contrary to the claim of Theorem 4.3 of [2]) Algorithm 2 need not converge to the values in concurrent safety games. However, as discussed before Algorithm 2 satisfies the monotonicity and optimality on termination, and for turn-based stochastic games (and also when restricted to $k$-uniform strategies) converges to the values as termination is guaranteed. In the example we will also argue how Algorithm 4 converges to the values of the game.

**Example 3** Our example consists of two steps. In the first step we will present a gadget where the value is irrational and with probability 1 absorbing states are reached.

**Step 1.** We first consider the game shown in Fig 3 with three states $\{s_0, s_1, s_2\}$ with two actions $a, b$ for player 1 and $c, d$ for player 2. The states $s_0, s_1$ are safe states, and $s_2$ is a non-safe state. The transitions are as follows: (1) $s_1$ and $s_2$ are absorbing; and (2) in $s_0$ we have the following transitions, (a) given action pairs $ac$ and $bd$ the next state is $s_1$, (b) given action pair $bc$ the next state is $s_2$, and (c) given action pair $ad$ the next states are $s_0$ and $s_1$ with probability $1/2$ each. In this game, the state $s_0$ is transient, as given any action pairs, the set $\{s_1, s_2\}$ of absorbing states is reached with probability at least $1/2$ in one step. Hence the set $\{s_1, s_2\}$ is reached with probability 1, irrespective of the choice of strategies of the players. Hence in this game the objective for player 1 is equivalently to reach $s_1$. Let us denote by $x$ the value of the game at $s_0$, and let us consider the following matrix

$$M = \begin{bmatrix} 1 & \frac{x}{2} \\ 0 & 1 \end{bmatrix}$$

Then $x = \min \max M$. In other words, consider the valuation $v_x = (x, 1, 0)$ for states $s_0, s_1$ and $s_2$, respectively, and $x = \min \max M$ describes that $v_x = Pre_1(v_x)$, and it is the least fixpoint of valuations

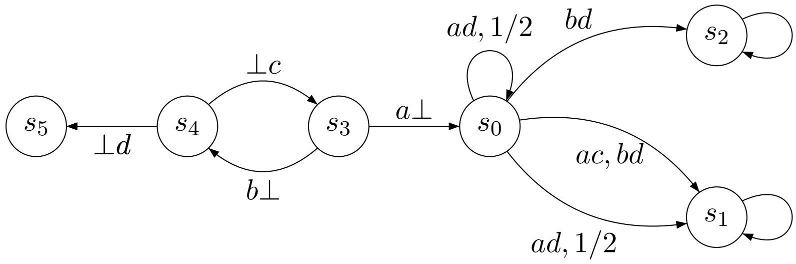


Figure 4: Counter example game.

satisfying $v = Pre_1(v)$. We now analyze the value $x$ at $s_0$. The solution of $x$ is achieved by solving the following optimization problem

$$\text{minimize } x \quad \text{subject to} \quad y + \big((1-y)\cdot x\big)/2 \le x \text{ and } 1-y \le x.$$

Intuitively, $y$ denotes the probability to choose move $a$ in an optimal strategy. The solution to the optimization problem is achieved by setting $x = 1-y$. Hence we have $y + (1-y)^2/2 = (1-y)$, i.e., $(1+y)^2 = 2$. Since $y$ must lie in the interval $[0,1]$, we have $y = \sqrt{2}-1$, and thus we have $x = 2-\sqrt{2} < 0.6$. We now analyze Algorithm 2 on this example. Let $v_i$ denote the valuation of the $i$-th iteration, and let $v_i^0$ be the value at state $s_0$. Then we have $v_i^0 < v_{i+1}^0$ and in the limit it converges to value $2-\sqrt{2}$. We observe that on this example Algorithm 2 exactly behaves as Algorithm 1 (strategy improvement for reachability) as the objective for player 1 is equivalently to reach $s_1$, since $s_0$ is transient. The reason of the strict inequality $v_i^0 < v_{i+1}^0$ is as follows: if the valuation at state $s_0$ in $i$-th and $i+1$-th iteration is the same, then by correctness of Algorithm 1 it follows that the values would have been achieved in finitely many steps, implying convergence to a rational value at $s_0$. The convergence to the values in the limit is due to correctness of Algorithm 1.

**Step 2.** We will now augment the game of Step 1 to construct an example to show that Algorithm 2 does not necessarily converge to the values. Consider the game shown in Fig 4 augmenting the game of Fig 3 with some additional states (states $s_3, s_4$ and $s_5$) and transitions (we only show the interesting transitions in the figure for simplicity). All the additional states shown are safe states. The value of state $s_5$ is $0.6$ (consider it as a probabilistic state going to state $s_1$ with probability $0.6$ and $s_2$ with probability $0.4$, and these edges are not shown in the figure). The transitions at state $s_3$ and $s_4$ are as follows: in state $s_3$, player 1 can goto state $s_0$ or $s_4$ by choosing actions $a$ and $b$, respectively (at $s_3$ player 2 has only one action $\bot$); and in state $s_4$, player 2 can goto state $s_3$ or $s_5$ by choosing actions $c$ and $d$, respectively (at $s_4$ player 1 has only one action $\bot$). We analyze Algorithm 2 on the example shown in Fig 4. In this game, at $s_3$ player 1 starts by playing actions $a$ and $b$ uniformly, and player 2 responds by chosing action $c$. In the iterations of the algorithm it follows by the argument of Step 1, that the set $I$ of Step 3.1 of Algorithm 2 is always non-empty as $s_0 \in I$. Hence in every iteration the value at $s_0$ improves, and the strategy in $s_3$ and $s_4$ does not change. Hence the valuation at $s_3$ converges to the valuation at $s_0$, i.e., to $2-\sqrt{2} < 0.6$. However, by switching to action $b$ at $s_3$, player 1 can enforce player 2 to play action $d$ at $s_4$ and ensure value $0.6$. In other words, the value at $s_3$ is $0.6$, whereas Algorithm 2 converges to $2-\sqrt{2} < 0.6$.

The switching to action $b$ would have been ensured by the turn-based construction of Step 3.3. For turn-based stochastic games or $k$-uniform memoryless strategies, since convergence to values is guaranteed, the turn-based construction of Step 3.3 is also ensured to get executed. However, as the convergence to values in concurrent games is in the limit, Step 3.3 of Algorithm 2 may not get executed as shown by this example. However, we now illustrate that the valuations of Algorithm 4 converges to the values. We consider Algorithm 4: Consider $k$-uniform strategies, for a finite $k \geq 2$, then the value at $s_0$ for $k$-uniform strategies converges in finitely many steps to a value smaller than $0.6$ (as it converges to a value smaller than the value at $s_0$), and then Step 3.3 of Algorithm 3 would get executed, and the value at $s_3$ would be assigned to $0.6$. In other words, for Algorithm 4 the values at $s_3, s_4$ and $s_5$ are always set to $0.6$, and the value at $s_0$ converges in the limit to $2 - \sqrt{2}$. Thus with the example we show that though Algorithm 2 does not necessarily converge to the values, Algorithm 4 correctly converges to the values. ∎

**Complexity.** Algorithm 2 may not terminate in general; we briefly describe the complexity of every iteration. Given a valuation $v_i$, the computation of $Pre_1(v_i)$ involves the solution of matrix games with rewards $v_i$; this can be done in polynomial time using linear programming. Given $v_i$, if $Pre_1(v_i) = v_i$, the sets $\mathsf{OptSel}(v_i, s)$ and $\mathsf{OptSelCount}(v_i, s)$ can be computed by enumerating the subsets of available actions at $s$ and then using linear-programming. For example, to check whether $(A, B) \in \mathsf{OptSelCount}(v_i, s)$ it suffices to check both of these facts:

1. *($A$ is the support of an optimal selector $\xi_1$).* there is an selector $\xi_1$ such that (i) $\xi_1$ is optimal (i.e. for all actions $b \in \Gamma_2(s)$ we have $Pre_{\xi_1,b}(v_i)(s) \geq v_i(s)$); (ii) for all $a \in A$ we have $\xi_1(a) > 0$, and for all $a \notin A$ we have $\xi_1(a) = 0$;

2. *($B$ is the set of counter-optimal actions against $\xi_1$).* for all $b \in B$ we have $Pre_{\xi_1,b}(v_i)(s) = v_i(s)$, and for all $b \notin B$ we have $Pre_{\xi_1,b}(v_i)(s) > v_i(s)$.

All the above checks can be performed by checking feasibility of sets of linear equalities and inequalities. Hence, $\mathsf{TB}(G, v_i, F)$ can be computed in time polynomial in size of $G$ and $v_i$ and exponential in the number of moves. We observe that the construction is exponential only in the number of moves at a state, and not in the number of states. The number of moves at a state is typically much smaller than the size of the state space. We also observe that the improvement step 3.3.2 requires the computation of the set of almost-sure winning states of a turn-based stochastic safety game: this can be done both via linear-time discrete graph-theoretic algorithms [4], and via symbolic algorithms [10]. Both of these methods are more efficient than the basic step 3.4 of the improvement algorithm, where the quantitative values of an MDP must be computed. Thus, the improvement step 3.3 of Algorithm 2 is in practice should not be inefficient, compared with the standard improvement steps 3.2 and 3.4. We now discuss the above steps for Algorithm 3. The argument is similar as above, but in case of $k$-uniform selectors, we need to ensure that the witness selectors are $k$-uniform which can be achieved with integer constraints. In other words, for Algorithm 3 the above checks are performed by checking feasibility of sets of integer linear equalities and inequalities (which can be achieved in exponential time). Again, the construction is exponential in the number of moves at a state, and not in the number of states. Hence we enumerate over sets of moves at a state (exponential in number of moves), and then need to solve integer linear constraints (the size of the integer linear constraints is polynomial in the number of moves, and is achieved in time exponential in the number of moves). Thus again the improvement step 3.3 of Algorithm 3 is polynomial in the size of the game, and exponential in the number of moves.

## 7.2 Termination for Approximation

In this subsection we present termination criteria for strategy improvement algorithms for concurrent games for $\varepsilon$-approximation.

**Termination for concurrent games.** We apply the reachability strategy improvement algorithm (Algorithm 1) for player 2, for a reachability objective $\mathrm{Reach}(T)$, we obtain a sequence of valuations $(u_i)_{i \geq 0}$ such that (a) $u_{i+1} \geq u_i$; (b) if $u_{i+1} = u_i$, then $u_i = \langle\langle 2 \rangle\rangle_{\mathsf{val}}(\text{Reach}(T))$; and (c) $\lim_{i\to\infty} u_i = \langle\langle 2 \rangle\rangle_{\mathsf{val}}(\text{Reach}(T))$. Given a concurrent game $G$ with $F \subseteq S$ and $T = S \setminus F$, we apply Algorithm 1 to obtain the sequence of valuation $(u_i)_{i\geq 0}$ as above, and we apply Algorithm 4 to obtain a sequence of valuation $(v_i)_{i\geq 0}$. The termination criteria are as follows:

1. if for some $i$ we have $u_{i+1} = u_i$, then we have $u_i = \langle\langle 2 \rangle\rangle_{\mathsf{val}}(\text{Reach}(T))$, and $1 - u_i = \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\text{Safe}(F))$, and we obtain the values of the game;
2. if for some $i$ we have $v_{i+1} = v_i$, then we have $1-v_i = \langle\langle 2 \rangle\rangle_{\mathsf{val}}(\text{Reach}(T))$, and $v_i = \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\text{Safe}(F))$, and we obtain the values of the game; and
3. for $\varepsilon > 0$, if for some $i \geq 0$, we have $u_i + v_i \geq 1 - \varepsilon$, then for all $s \in S$ we have $v_i(s) \geq \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\text{Safe}(F))(s) - \varepsilon$ and $u_i(s) \geq \langle\langle 2 \rangle\rangle_{\mathsf{val}}(\text{Reach}(T))(s) - \varepsilon$ (i.e., the algorithm can stop for $\varepsilon$-approximation).

Observe that since $(u_i)_{i\geq 0}$ and $(v_i)_{i\geq 0}$ are both monotonically non-decreasing and $\langle\langle 1 \rangle\rangle_{\mathsf{val}}(\text{Safe}(F)) + \langle\langle 2 \rangle\rangle_{\mathsf{val}}(\text{Reach}(T)) = 1$, it follows that if $u_i + v_i \geq 1 - \varepsilon$, then forall $j \geq i$ we have $u_i \geq u_j - \varepsilon$ and $v_i \geq v_j - \varepsilon$. This establishes that $u_i \geq \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\text{Safe}(F)) - \varepsilon$ and $v_i \geq \langle\langle 2 \rangle\rangle_{\mathsf{val}}(\text{Reach}(T)) - \varepsilon$; and the correctness of the stopping criteria (3) for $\varepsilon$-approximation follows. We also note that instead of applying the reachability strategy improvement algorithm, a value-iteration algorithm can be applied for reachability games to obtain a sequence of valuation with properties similar to $(u_i)_{i\geq 0}$ and the above termination criteria can be applied.

**Theorem 9** *Let $G$ be a concurrent game structure with a safety objective Safe$(F)$. Algorithm 4 and Algorithm 1 for player 2 for the reachability objective Reach$(S \setminus F)$ yield two sequences of monotonic valuations $(v_i)_{i\geq 0}$ and $(u_i)_{i\geq 0}$, respectively, such that (a) for all $i \geq 0$, we have $v_i \leq \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\mathit{Safe}(F)) \leq 1 - u_i$; and (b) $\lim_{i\to\infty} v_i = \lim_{i\to\infty} 1 - u_i = \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\mathit{Safe}(F))$.*

**Bounds for approximation.** We now discuss the bounds for approximation for concurrent games with reachability objectives, which follows from the results of [18, 17]. It follows from the results of [18] that for all $\varepsilon > 0$, there exist $k$-uniform memoryless optimal strategies for concurrent reachability and safety games $G$, where $k$ is bounded by $(\frac{1}{\varepsilon})^{2^{O(|G|)}}$. It follows that for all $\varepsilon > 0$, if we consider our strategy improvement algorithm (restricted to $k$-uniform selectors) for reachability games, then upon termination the valuation obtained is an $\varepsilon$-approximation of the value function of the game, where $k$ is bounded by $(\frac{1}{\varepsilon})^{2^{O(|G|)}}$. Using the restriction to $k$-uniform memoryless strategies, along with the reduction of concurrent games to turn-based stochastic game for $k$-uniform memoryless strategies and the termination bound for turn-based stochastic games we obtain a double exponential bound on the number of iterations required for termination (note that if $k = (\frac{1}{\varepsilon})^{2^{O(|G|)}}$, then the total number of $k$-uniform memoryless strategies is $k^{O(|G|)}$, which is double exponential) (also see [17] for details). Moreover, the recent result of [17] shows that the double exponential bound is near optimal for the strategy improvment algorithm for concurrent games with reachability objectives.

**Approximation of strategies.** The previous method to solve concurrent reachability and safety games was the value-iteration algorithm. The witness strategy produced by the value-iteration algorithm for concurrent reachability games is not memoryless; and for concurrent safety games since the value-iteration algorithm converges from above it does not provide any witness strategies. The only previous algorithm to approximate memoryless $\varepsilon$-optimal strategies, for $\varepsilon > 0$, for concurrent reachability and safety games is the naive algorithm that exhaustively searches over the set of all $k$-uniform memoryless strategies (such that the $k$-uniform memoryless strategies suffices for $\varepsilon$-optimality and $k$-depends in $\varepsilon$). Our strategy improvement algorithms for concurrent reachability and safety games are the first strategy search based approach to approximate $\varepsilon$-optimal strategies.

**Acknowledgements.** We thanks anonymous reviewers for many helpful and insightful comments that significantly improved the presentation of the paper, and help us fix gaps in the results of the conference versions. We warmly acknowlede their help. This work was partially supported by ERC Start Grant Graph Games (Project No 279307), FWF NFN Grant S11407-N23 (RiSE) and a Microsoft faculty fellowship.

# References

[1] D.P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 1995. Volumes I and II.

[2] K. Chatterjee, L. de Alfaro, and T. A. Henzinger. Termination criteria for solving concurrent safety and reachability games. In *SODA*, pages 197–206. ACM-SIAM, 2009.

[3] K. Chatterjee, L. de Alfaro, and T.A. Henzinger. Strategy improvement in concurrent reachability games. In *QEST'06*. IEEE, 2006.

[4] K. Chatterjee, M. Jurdziński, and T.A. Henzinger. Simple stochastic parity games. In *CSL'03*, volume 2803 of *LNCS*, pages 100–113. Springer, 2003.

[5] A. Condon. The complexity of stochastic games. *Information and Computation*, 96(2):203–224, 1992.

[6] A. Condon. On algorithms for simple stochastic games. In *Advances in Computational Complexity Theory*, volume 13 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, pages 51–73. American Mathematical Society, 1993.

[7] C. Courcoubetis and M. Yannakakis. The complexity of probabilistic verification. *Journal of the ACM*, 42(4):857–907, 1995.

[8] L. de Alfaro. *Formal Verification of Probabilistic Systems*. PhD thesis, Stanford University, 1997. Technical Report STAN-CS-TR-98-1601.

[9] L. de Alfaro and T.A. Henzinger. Concurrent omega-regular games. In *Proceedings of the 15th Annual Symposium on Logic in Computer Science*, pages 141–154. IEEE Computer Society Press, 2000.

[10] L. de Alfaro, T.A. Henzinger, and O. Kupferman. Concurrent reachability games. *Theoretical Computer Science*, 386(3):188–217, 2007.

[11] L. de Alfaro and R. Majumdar. Quantitative solution of omega-regular games. *Journal of Computer and System Sciences*, 68:374–397, 2004.

[12] C. Derman. *Finite State Markovian Decision Processes*. Academic Press, 1970.

[13] K. Etessami and M. Yannakakis. Recursive concurrent stochastic games. In *ICALP 06: Automata, Languages, and Programming*. Springer, 2006.

[14] H. Everett. Recursive games. In *Contributions to the Theory of Games III*, volume 39 of *Annals of Mathematical Studies*, pages 47–78, 1957.

[15] J. Filar and K. Vrieze. *Competitive Markov Decision Processes*. Springer-Verlag, 1997.

[16] H. Gimbert and F. Horn. Simple stochastic games with few random vertices are easy to solve. In *FoSSaCS'08*, 2008.

[17] K. A. Hansen, R. Ibsen-Jensen, and P. B. Miltersen. The complexity of solving reachability games using value and strategy iteration. In *CSR*, pages 77–90, 2011.

[18] K. A. Hansen, M. Koucký, and P. . Miltersen. Winning concurrent reachability games requires doubly-exponential patience. In *LICS*, pages 332–341, 2009.

[19] A.J. Hoffman and R.M. Karp. On nonterminating stochastic games. *Management Sciences*, 12(5):359–370, 1966.

[20] R. A. Howard. *Dynamic Programming and Markov Processes*. MIT Press, 1960.

[21] J.G. Kemeny, J.L. Snell, and A.W. Knapp. *Denumerable Markov Chains*. D. Van Nostrand Company, 1966.

[22] T. A. Liggett and S. A. Lippman. Stochastic games with perfect information and time average payoff. *Siam Review*, 11:604–607, 1969.

[23] J.F. Mertens and A. Neyman. Stochastic games. *International Journal of Game Theory*, 10:53–66, 1981.

[24] T. Parthasarathy. Discounted and positive stochastic games. *Bulletin of American Mathematical Society*, 77(1):134–136, 1971.

[25] S. S. Rao, R. Chandrasekaran, and K. P. K. Nair. Algorithms for discounted games. *Journal of Opt. Theory and Applications*, pages 627–637, 1973.

[26] L.S. Shapley. Stochastic games. *Proc. Nat. Acad. Sci. USA*, 39:1095–1100, 1953.

[27] E. Solan. Continuity of the value of competitive Markov decision processes. *Journal of Theoretical Probability*, 16:831–845, 2003.

[28] U. Zwick and M.S. Paterson. The complexity of mean payoff games on graphs. *Theoretical Computer Science*, 158:343–359, 1996.

**Algorithm 3** $k$-Uniform Restricted Safety Strategy-Improvement Algorithm

**Input:** a concurrent game structure $G$ with safe set $F$, and number $k$.
**Output:** a strategy $\overline{\gamma}$ for player 1.
0. Compute $W_1 = \{s \in S \mid \langle\langle 1 \rangle\rangle_{\mathsf{val}}(\mathsf{Safe}(F))(s) = 1\}$; and $k = \max\{k, |M|\}$.
1. Let $\gamma_0 = \xi_1^{\mathit{unif}}$ and $i = 0$.
2. Compute $v_0 = \langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_0}(\mathsf{Safe}(F))$.
3. **do** $\{$
   3.1. Let $I_k = \{s \in S \setminus (W_1 \cup T) \mid \sup_{\xi_1' \in \Lambda^k(s)} Pre_{1:\xi_1'}(v_i)(s) > v_i(s)\}$.
   3.2 **if** $I_k \neq \emptyset$, **then**
      3.2.1 Let $\xi_1$ be a $k$-uniform player-1 selector such that for all states $s \in I$,
         we have $Pre_{1:\xi_1}(v_i)(s) = \sup_{\xi_1' \in \Lambda^k(s)} Pre_{1:\xi_1'}(v_i)(s) > v_i(s)$.
      3.2.2 The player-1 selector $\gamma_{i+1}$ is defined as follows: for each state $s \in S$, let

$$\gamma_{i+1}(s) = \begin{cases} \gamma_i(s) & \text{if } s \notin I_k; \\ \xi_1(s) & \text{if } s \in I_k. \end{cases}$$

   3.3 **else**
      3.3.1 let$(\overline{G}_{v_i}^k, \overline{F}) = \mathsf{TB}(G, v_i, F, k)$
      3.3.2 let $\overline{A}_i^k$ be the set of almost-sure winning states in $\overline{G}_{v_i}^k$ for $\mathsf{Safe}(\overline{F})$ and
         $\overline{\pi}_1$ be a pure memoryless almost-sure winning strategy from the set $\overline{A}_i^k$.
      3.3.3 **if** $((\overline{A}_i^k \cap S) \setminus W_1 \neq \emptyset)$
         3.3.3.1 let $U = (\overline{A}_i^k \cap S) \setminus W_1$
         3.3.3.2 The player-1 selector $\gamma_{i+1}$ is defined as follows: for $s \in S$, let

$$\gamma_{i+1}(s) = \begin{cases} \gamma_i(s) & \text{if } s \notin U; \\ \xi_1(s) & \text{if } s \in U, \xi_1(s) \in \mathsf{OptSel}(v_i, s, k), \\ & \quad \overline{\pi}_1(s) = (s, A, B), B = \mathsf{OptSelCount}(s, v, \xi_1, k). \end{cases}$$

   3.4. Compute $v_{i+1} = \langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_{i+1}}(\mathsf{Safe}(F))$.
   3.5. Let $i = i + 1$.
$\}$ **until** $I_k = \emptyset$ and $(\overline{A}_{i-1}^k \cap S) \setminus W_1 = \emptyset$.
4. **return** $\overline{\gamma}_i$.

**Algorithm 4** Convergent Safety Strategy-Improvement Algorithm

**Input:** a concurrent game structure $G$ with safe set $F$.
**Output:** a strategy $\overline{\gamma}$ for player 1.
0. $k = |M|$ and $i = 0$.
1. **do** $\{$
 1.1 $\gamma_{i+1} =$ Algorithm 3($G, F, k$)
 1.2 Compute $v_{i+1} = \langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\overline{\gamma}_{i+1}}(\mathsf{Safe}(F))$
 1.3 Let $I = \{s \in S \setminus (W_1 \cup T) \mid Pre_1(v_i)(s) > v_i(s)\}$.
 1.4 Let$(\overline{G}_{v_i}, \overline{F}) = \mathsf{TB}(G, v_i, F)$
  1.4.1 let $\overline{A}_i$ be the set of almost-sure winning states in $\overline{G}_{v_i}$ for Safe$(\overline{F})$.
 1.5 Let $i = i + 1$ and $k = k + 1$.
$\}$ **until** $I = \emptyset$ and $(\overline{A}_{i-1} \cap S) \setminus W_1 = \emptyset$.
2. **return** $\overline{\gamma}_i$.

# 8 Technical Appendix

We now present the details of restriction to $k$-uniform selectors, and the details of the notations used in Algorithm 3. The definitions are essentially same as for selectors and optimal selectors, but restricted to $k$-uniform selectors.

**Optimal $k$-uniform selectors.** For $k > 0$, a valuation $v$ and a state $s$, let

$$Pre_1^k(v)(s) = \sup_{\xi_1' \in \Lambda_1^k(s)} Pre_{1:\xi_1'}(v)(s).$$

denote the optimal one-step value among $k$-uniform selectors. For $k > 0$, given a valuation $v$ and a state $s$, we define by

$$\mathsf{OptSel}(v, s, k) = \{\xi_1 \in \Lambda_1^k(s) \mid Pre_{1:\xi_1}(v)(s) = Pre_1^k(v)(s)\}$$

the set of optimal selectors among $k$-uniform selectors for $v$ at state $s$. For a $k$-uniform optimal selector $\xi_1 \in \mathsf{OptSel}(v, s, k)$, we define the set of counter-optimal actions as follows:

$$\mathsf{CountOpt}(v, s, \xi_1, k) = \{b \in \Gamma_2(s) \mid Pre_{\xi_1,b}(v)(s) = Pre_1^k(v)(s)\}.$$

Observe that for $\xi_1 \in \mathsf{OptSel}(v, s, k)$, for all $b \in \Gamma_2(s) \setminus \mathsf{CountOpt}(v, s, \xi_1, k)$ we have $Pre_{\xi_1,b}(v)(s) > Pre_1^k(v)(s)$. We define the set of $k$-uniform optimal selector support and the counter-optimal action set as follows:

$$\begin{aligned} \mathsf{OptSelCount}(v, s, k) \;=\; & \{(A, B) \subseteq \Gamma_1(s) \times \Gamma_2(s) \mid \exists \xi_1 \in \Lambda_1^k(s).\ \xi_1 \in \mathsf{OptSel}(v, s, k) \\ & \wedge\ Supp(\xi_1) = A\ \wedge\ \mathsf{CountOpt}(v, s, \xi_1, k) = B\}; \end{aligned}$$

i.e., it consists of pairs $(A, B)$ of actions of player 1 and player 2, such that there is a $k$-uniform optimal selector $\xi_1$ with support $A$, and $B$ is the set of counter-optimal actions to $\xi_1$.

**Turn-based reduction.** Given a concurrent game $G = \langle S, M, \Gamma_1, \Gamma_2, \delta\rangle$, a valuation $v$, and bound $k$ for $k$-uniformity we construct a turn-based stochastic game $\overline{G}_v^k = \langle (\overline{S}, \overline{E}), (\overline{S}_1, \overline{S}_2, \overline{S}_R), \overline{\delta}\rangle$ as follows:

1. The set of states is as follows:

$$\begin{aligned} \overline{S} \;=\; & S \cup \{(s, A, B) \mid s \in S,\ (A, B) \in \mathsf{OptSelCount}(v, s, k)\} \\ \cup\; & \{(s, A, b) \mid s \in S,\ (A, B) \in \mathsf{OptSelCount}(v, s, k),\ b \in B\}. \end{aligned}$$

2. The state space partition is as follows: $\overline{S}_1 = S$; $\overline{S}_2 = \{(s, A, B) \mid s \in S, (A, B) \in \mathsf{OptSelCount}(v, s, k)\}$; and $\overline{S}_R = \{(s, A, b) \mid s \in S,\ (A, B) \in \mathsf{OptSelCount}(v, s, k), b \in B\}$. In other words, $(\overline{S}_1, \overline{S}_2, \overline{S}_R)$ is a partition of the state space, where $\overline{S}_1$ are player 1 states, $\overline{S}_2$ are player 2 states, and $\overline{S}_R$ are random or probabilistic states.

3. The set of edges is as follows:

$$\begin{aligned} \overline{E} \;=\; & \{(s, (s, A, B)) \mid s \in S, (A, B) \in \mathsf{OptSelCount}(v, s, k)\} \\ \cup\; & \{((s, A, B), (s, A, b)) \mid b \in B\} \cup \{((s, A, b), t) \mid t \in \bigcup_{a \in A} Dest(s, a, b)\}. \end{aligned}$$

4. The transition function $\overline{\delta}$ for all states in $\overline{S}_R$ is uniform over its successors.

Intuitively, the reduction is as follows. Given the valuation $v$, state $s$ is a player 1 state where player 1 can select a pair $(A, B)$ (and move to state $(s, A, B)$) with $A \subseteq \Gamma_1(s)$ and $B \subseteq \Gamma_2(s)$ such that there is a $k$-uniform optimal selector $\xi_1$ with support exactly $A$ and the set of counter-optimal actions to $\xi_1$ is the set $B$. From a player 2 state $(s, A, B)$, player 2 can choose any action $b$ from the set $B$, and move to state $(s, A, b)$. A state $(s, A, b)$ is a probabilistic state where all the states in $\bigcup_{a \in A} Dest(s, a, b)$ are chosen uniformly at random. Given a set $F \subseteq S$ we denote by $\overline{F} = F \cup \{(s, A, B) \in \overline{S} \mid s \in F\} \cup \{(s, A, b) \in \overline{S} \mid s \in F\}$. We refer to the above reduction as $\mathsf{TB}$, i.e., $(\overline{G}_v^k, \overline{F}) = \mathsf{TB}(G, v, F, k)$.

**Proof. (of Lemma 17).** The proof of the result is essentially identical as the proof of Theorem 7, and we present the details for completeness. Let $v_i = z_i^k$. We show that for all $k$-uniform memoryless strategies $\pi_1$ for player 1 we have $\langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\pi_1}(\mathsf{Safe}(F)) \leq v_i$.

Let $\overline{\pi}_2$ be a pure memoryless optimal strategy for player 2 in $\overline{G}_{v_i}^k$ for the objective complementary to $\mathsf{Safe}(\overline{F})$, where $(\overline{G}_{v_i}^k, \mathsf{Safe}(\overline{F})) = \mathsf{TB}(G, v_i, F, k)$. Consider a $k$-uniform memoryless strategy $\pi_1$ for player 1, and we define a pure memoryless strategy $\pi_2$ for player 2 as follows.

1. If $\pi_1(s) \notin \mathsf{OptSel}(v_i, s, k)$, then $\pi_2(s) = b \in \Gamma_2(s)$, such that $Pre_{\pi_1(s), b}(v_i)(s) < v_i(s)$; (such a $b$ exists since $\pi_1(s) \notin \mathsf{OptSel}(v_i, s, k)$).

2. If $\pi_1(s) \in \mathsf{OptSel}(v_i, s, k)$, then let $A = Supp(\pi_1(s))$, and consider $B$ such that $B = \mathsf{OptSelCount}(v_i, s, \pi_1(s), k)$. Then we have $\pi_2(s) = b$, such that $\overline{\pi}_2((s, A, B)) = (s, A, b)$.

Observe that by construction of $\pi_2$, for all $s \in S \setminus (W_1 \cup T)$, we have $Pre_{\pi_1(s), \pi_2(s)}(v_i)(s) \leq v_i(s)$. We first show that in the Markov chain obtained by fixing $\pi_1$ and $\pi_2$ in $G$, there is no closed connected recurrent set of states $C$ such that $C \subseteq S \setminus (W_1 \cup T)$. Assume towards contradiction that $C$ is a closed connected recurrent set of states in $S \setminus (W_1 \cup T)$. The following case analysis achieves the contradiction.

1. Suppose for every state $s \in C$ we have $\pi_1(s) \in \mathsf{OptSel}(v_i, s, k)$. Then consider the strategy $\overline{\pi}_1$ in $\overline{G}_{v_i}^k$ such that for a state $s \in C$ we have $\overline{\pi}_1(s) = (s, A, B)$, where $\pi_1(s) = A$, and $B = \mathsf{OptSelCount}(v_i, s, \pi_1(s), k)$. Since $C$ is closed connected recurrent states, it follows by construction that for all states $s \in C$ in the game $\overline{G}_{v_i}^k$ we have $\Pr_s^{\overline{\pi}_1, \overline{\pi}_2}(\mathsf{Safe}(\overline{C})) = 1$, where $\overline{C} = C \cup \{(s, A, B) \mid s \in C\} \cup \{(s, A, b) \mid s \in C\}$. It follows that for all $s \in C$ in $\overline{G}_{v_i}^k$ we have $\Pr_s^{\overline{\pi}_1, \overline{\pi}_2}(\mathsf{Safe}(\overline{F})) = 1$. Since $\overline{\pi}_2$ is an optimal strategy, it follows that $C \subseteq (\overline{A}_i^k \cap S) \setminus W_1$. This contradicts that $(\overline{A}_i^k \cap S) \setminus W_1 = \emptyset$.

2. Otherwise for some state $s^* \in C$ we have $\pi_1(s^*) \notin \mathsf{OptSel}(v_i, s^*, k)$. Let $r = \min\{q \mid U_q(v_i) \cap C \neq \emptyset\}$, i.e., $r$ is the least value-class with non-empty intersection with $C$. Hence it follows that for all $q < r$, we have $U_q(v_i) \cap C = \emptyset$. Observe that since for all $s \in C$ we have $Pre_{\pi_1(s), \pi_2(s)}(v_i)(s) \leq v_i(s)$, it follows that for all $s \in U_r(v_i)$ either (a) $Dest(s, \pi_1(s), \pi_2(s)) \subseteq U_r(v_i)$; or (b) $Dest(s, \pi_1(s), \pi_2(s)) \cap U_q(v_i) \neq \emptyset$, for some $q < r$. Since $U_r(v_i)$ is the least value-class with non-empty intersection with $C$, it follows that for all $s \in U_r(v_i)$ we have $Dest(s, \pi_1(s), \pi_2(s)) \subseteq U_r(v_i)$. It follows that $C \subseteq U_r(v_i)$. Consider the state $s^* \in C$ such that $\pi_1(s^*) \notin \mathsf{OptSel}(v_i, s, k)$. By the construction of $\pi_2(s)$, we have $Pre_{\pi_1(s^*), \pi_2(s^*)}(v_i)(s^*) < v_i(s^*)$. Hence we must have $Dest(s^*, \pi_1(s^*), \pi_2(s^*)) \cap U_q(v_i) \neq \emptyset$, for some $q < r$. Thus we have a contradiction.

It follows from above that there is no closed connected recurrent set of states in $S \setminus (W_1 \cup T)$, and hence with probability 1 the game reaches $W_1 \cup T$ from all states in $S \setminus (W_1 \cup T)$. Hence the probability to satisfy $\mathsf{Safe}(F)$ is equal to the probability to reach $W_1$. Since for all states $s \in S \setminus (W_1 \cup T)$ we have

$Pre_{\pi_1(s),\pi_2(s)}(v_i)(s) \leq v_i(s)$, it follows that given the strategies $\pi_1$ and $\pi_2$, the valuation $v_i$ satisfies all the inequalities for linear program to reach $W_1$. It follows that the probability to reach $W_1$ from $s$ is atmost $v_i(s)$. It follows that for all $s \in S \setminus (W_1 \cup T)$ we have $\langle\langle 1 \rangle\rangle_{\mathsf{val}}^{\pi_1}(\mathsf{Safe}(F))(s) \leq v_i(s)$. This completes the proof. ∎