View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by IST Austria: PubRep (Institute of Science and Technology)

Discounting in Games across Time Scales*

Krishnendu Chatterjee
IST Austria (Institute of Science and Technology Austria)
krishnendu.chatterjee@ist.ac.at

Rupak Majumdar
University of California, Los Angeles, USA and
Max Planck Institute for Software Systems, Germany
rupak@cs.ucla.edu

We introduce two-level discounted games played by two players on a perfect-information stochastic game graph. The upper level game is a discounted game and the lower level game is an undiscounted reachability game. Two-level games model hierarchical and sequential decision making under uncertainty across different time scales. We show the existence of pure memoryless optimal strategies for both players and an ordered field property for such games. We show that if there is only one player (Markov decision processes), then the values can be computed in polynomial time. It follows that whether the value of a player is equal to a given rational constant in two-level discounted games can be decided in NP $\cap$ coNP. We also give an alternate strategy improvement algorithm to compute the value.

1 Introduction

Discrete stochastic games have been extensively studied as models for decision making under adversarial interactions in an uncertain environment, and have found many applications, such as in manufacturing systems control and inventory management [4].

In many such applications, the interaction with the environment occurs in a hierarchical manner, intercalated across different time scales. In the short-term, a decision has to be made about choosing one of several possible actions. For example, short term decisions can determine whether to buy a certain product or another, or whether to increase or decrease production capacity. In the long-term, the system gets a profit or a loss at each step based on its existing inventory. Both short-term and long-term decisions can potentially involve uncertainty and adversarial interactions. Moreover, long term decisions are influenced by the short term actions chosen, and model the effect of the local decisions on the overall profits or losses of the system.

Technically, the two types of interaction are modeled using two distinct classes of games. Undiscounted *reachability games* are used to model short-term decision making (e.g., what action to take next). In a reachability game on a state space, one fixes a set of goal states, and the objective of player 1 is to maximize the probability of reaching the goal states.

On the other hand, *discounted reward games* model long-term rewards for the system (e.g., how the actions chosen locally relate to long-term profits). In a discounted game, player 1 gets a reward in each step, and a time discount parameter $\lambda \in (0,1)$ is used to "discount" the reward at future time points (i.e., the reward r obtained t time units in the future is given a value $\lambda^t r$). The short-term interactions are abstracted away into an atomic step that uniformly sets the time granularity. The objective of player 1 is to maximize the expected normalized sum of discounted rewards.

To make the games concrete, consider economic policymakers setting financial policy. The specific policy implemented (e.g., the interest rate or the amount of regulation) affects the long-term health of the

*This research was funded in part by the US National Science Foundation grants CCF-0546170 and CNS-0702881, and DARPA grant HR0011-09-1-0037.

Angelo Montanari, Margherita Napoli, Mimmo Parente (Eds.)
Proceedings of GandALF 2010
EPTCS 25, 2010, pp. 22–29, doi:10.4204/EPTCS.25.6

© K.Chatterjee and R. Majumdar
This work is licensed under the
Creative Commons Attribution License.

economy, and the interplay between financial policy and the market can be modeled using discounted rewards. However, in each step, the specific policy chosen depends on "short-term" games between various stakeholders, such as politicians, the treasury, companies, and various interest groups. In this setting, the time granularity of long-term steps (policy implementation) is variable, and depends on the length and outcome of the short-term steps (deciding which policy to implement).

While each game model in itself provides a sound theoretical basis for reasoning about system behavior, the hierarchical interaction and varying time granularities (short-term vs. long-term) present in many applications is not adequately captured by either model. In this paper, we introduce models for such multi-level interactions and algorithms for sequential decision making in a setting where the time granularity can be variable. We define a *two-level* discounted game, in which a "lower level" reachability game is used to decide actions for a "higher level" discounted game. The discount factor is applied to the time scale of the higher-level game, not for every step that elapses in the lower-level game. Since every lower level game is different we obtain complete independence in granularity of transitions.

Our main result is the existence of value and pure memoryless strategies in two-level discounted games. Moreover, we show that the value and optimal strategies can be computed in polynomial time for Markov decision processes, and the complexity of checking if the value is equal to a rational is in NP $\cap$ co-NP for $2\frac{1}{2}$-player two-level discounted games. Two-level discounted games subsume classical discounted games, and our complexity bounds match the best known results for classical discounted games.

Technically, we combine the existence of pure memoryless strategies in discounted games [7] with the existence of pure memoryless strategies in (undiscounted) $2\frac{1}{2}$-player reachability games [5, 4, 6] together with a reduction from two-level games to a one-level discounted game. In particular, we show that for Markov decision processes, we can formulate the value at a state as a linear programming problem over the states of the two-level game. Thus, the games have an ordered field property: if all constants in the definition of the game come from a field F, then the value is also in F; in particular, games with rational probabilities and rational discount factors have a rational value. Together with the existence of pure memoryless strategies, this implies that the decision problem to check if the value is equal to a given rational is in NP $\cap$ co-NP. We also give a strategy improvement algorithm to compute the value, by combining strategy improvement algorithms for stochastic reachability [2] and discounted games [4].

Thus our new model of stochastic games provides a uniform framework for decision making across different time scales, and our algorithms show how to decide optimally in such a framework.

2 Definitions

We consider several classes of turn-based games: two-player turn-based probabilistic games ($2\frac{1}{2}$-player games), two-player turn-based deterministic games (2-player games), and Markov decision processes ($1\frac{1}{2}$-player games).

Notation. For a finite 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 $\mathscr{D}(A)$. Given a distribution $\delta \in \mathscr{D}(A)$, we denote by $\mathrm{Supp}(\delta) = \{x \in A \mid \delta(x) > 0\}$ the *support* of δ.

Game graphs. A *turn-based probabilistic game graph* ($2\frac{1}{2}$*-player game graph*) $G = ((S,E),(S_1,S_2,S_P),\delta)$ consists of a directed graph (S,E), a partition (S_1, S_2, S_P) of the finite set S of states, and a probabilistic transition function δ: $S_P \to \mathscr{D}(S)$, where $\mathscr{D}(S)$ 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_P are the *probabilistic* states, where the successor state is chosen according to the probabilistic transition function δ. We assume that for $s \in S_P$ 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. The size of a game graph $G = ((S,E),(S_1,S_2,S_P),\delta)$ is

$$|G| = |S| + |E| + \sum_{t \in S} \sum_{s \in S_P} |\delta(s)(t)|;$$

where $|\delta(s)(t)|$ denotes the space to represent the transition probability $\delta(s)(t)$ in binary.

The *turn-based deterministic game graphs* (*2-player game graphs*) are the special case of the $2\frac{1}{2}$-player game graphs with $S_P = \emptyset$. The *Markov decision processes* (*$1\frac{1}{2}$-player game graphs*) are the special case of the $2\frac{1}{2}$-player game graphs with $S_1 = \emptyset$ or $S_2 = \emptyset$. We refer to the MDPs with $S_2 = \emptyset$ as *player*-1 *MDPs*, and to the MDPs with $S_1 = \emptyset$ as *player*-2 *MDPs*.

Plays and strategies. An infinite path, or *play*, of the game graph G is an infinite sequence $\omega = \langle s_0, s_1, s_2, \ldots \rangle$ of states such that $(s_k, s_{k+1}) \in E$ for all $k \in \mathbb{N}$. We write Ω for the set of all plays, and for a state $s \in S$, we write $\Omega_s \subseteq \Omega$ for the set of plays that start from the state s.

A *strategy* for player 1 is a function $\sigma\colon S^* \cdot S_1 \to \mathscr{D}(S)$ that assigns a probability distribution to all finite sequences $\vec{w} \in S^* \cdot S_1$ of states ending in a player-1 state (the sequence represents a prefix of a play). Player 1 follows the strategy σ if in each player-1 move, given that the current history of the game is $\vec{w} \in S^* \cdot S_1$, she chooses the next state according to the probability distribution $\sigma(\vec{w})$. A strategy must prescribe only available moves, i.e., for all $\vec{w} \in S^*$, and $s \in S_1$ we have $\mathrm{Supp}(\sigma(\vec{w} \cdot s)) \subseteq E(s)$. The strategies for player 2 are defined analogously. We denote by Σ and Π the set of all strategies for player 1 and player 2, respectively.

Once a starting state $s \in S$ and strategies $\sigma \in \Sigma$ and $\pi \in \Pi$ for the two players are fixed, the outcome of the game is a random walk $\omega_s^{\sigma,\pi}$ for which the probabilities of events are uniquely defined, where an *event* $\mathscr{A} \subseteq \Omega$ is a measurable set of paths. For a state $s \in S$ and an event $\mathscr{A} \subseteq \Omega$, we write $\Pr_s^{\sigma,\pi}(\mathscr{A})$ for the probability that a path belongs to $\mathscr{A}$ if the game starts from the state s and the players follow the strategies σ and π, respectively. Similarly we denote by $\mathbb{E}_s^{\sigma,\pi}(\cdot)$ the expectation under the probability measure $\Pr_s^{\sigma,\pi}(\cdot)$. In the context of player-1 MDPs we often omit the argument π, because Π is a singleton set.

We classify strategies according to their use of randomization and memory. Strategies that do not use randomization are called pure; formally, a player-1 strategy σ is *pure* if for all $\vec{w} \in S^*$ and $s \in S_1$, there is a state $t \in S$ such that $\sigma(\vec{w} \cdot s)(t) = 1$. We denote by $\Sigma^P \subseteq \Sigma$ the set of pure strategies for player 1. In order to emphasize the potential use of randomization, we call a (general) strategy *randomized*. Let $\mathtt{M}$ be a set called *memory*, that is, $\mathtt{M}$ is a set of memory elements. A player-1 strategy σ can be described as a pair of functions $\sigma = (\sigma_u, \sigma_m)$: a *memory-update* function $\sigma_u\colon S \times \mathtt{M} \to \mathtt{M}$ and a *next-move* function $\sigma_m\colon S_1 \times \mathtt{M} \to \mathscr{D}(S)$. We can think of strategies with memory as input/output automata computing the strategies (see [3] for details). A strategy $\sigma = (\sigma_u, \sigma_m)$ is *finite-memory* if the memory $\mathtt{M}$ is finite, and then the size of the strategy σ, denoted as $|\sigma|$, is the size of its memory $\mathtt{M}$, i.e., $|\sigma| = |\mathtt{M}|$. We denote by Σ^F the set of finite-memory strategies for player 1, and by Σ^{PF} the set of *pure finite-memory* strategies; that is, $\Sigma^{PF} = \Sigma^P \cap \Sigma^F$. The strategy (σ_u, σ_m) is *memoryless* if $|\mathtt{M}| = 1$; that is, the next move does not depend on the history of the play but only on the current state. A memoryless player-1 strategy can be represented as a function $\sigma\colon S_1 \to \mathscr{D}(S)$. A *pure memoryless strategy* is a pure strategy that is memoryless. A pure memoryless strategy for player 1 can be represented as a function $\sigma\colon S_1 \to S$. We denote by Σ^M the set of memoryless strategies for player 1, and by Σ^{PM} the set of pure memoryless strategies; that is,

$\Sigma^{PM} = \Sigma^P \cap \Sigma^M$. Analogously we define the corresponding strategy families Π^P, Π^F, Π^{PF}, Π^M, and Π^{PM} for player 2.

Two-level discounted games. A *two-level discounted game* consists of a turn-based probabilistic game graph G; a partion of the state space S into (S_u, S_l) the set S_u of upper level states and the set S_l of lower level states; and a reward function $r : S_u \to \mathbb{R}_{>0}$ that maps every upper level state to a positive real-valued reward. We also require that from every state $s \in S_l$ player 1 can ensure to reach a state in S_u with probability 1. In other words, for all $s \in S_l$, there exists a player 1 strategy σ such that against all player 2 strategies π we have $\Pr_s^{\sigma,\pi}(\text{Reach}(S_u)) = 1$, where $\text{Reach}(S_u)$ is the set of paths that visit a state in S_u.

Discounted objectives. An objective f is a measurable function $f : \Omega \to \mathbb{R}$ that assigns to every path a real-valued payoff. The *discounted objective* in two-level discounted games is a measurable function $\text{TwoDisc} : \Omega \to \mathbb{R}$ defined as follows: for $0 < \beta < 1$, consider a path $\omega = \langle s_0, s_1, s_2, \ldots \rangle$ and for an index $i \geq 0$, let

$$\alpha(i) = \begin{cases} 0 & s_i \in S_l; \\ \beta^k \cdot r(s_i) & s_i \in S_u \text{ and the number of } S_u \text{ states in } \langle s_0, \ldots, s_{i-1} \rangle \text{ is } k-1; \end{cases}$$

then

$$\text{TwoDisc}(\omega) = (1-\beta) \cdot \sum_{i=0}^{\infty} \alpha(i).$$

In other words, the payoff of a path is the normalized discounted sum of the rewards of the path and the discounting is applied for every upper level state.

Optimal strategies. Given objectives f and $-f$ for player 1 and player 2, respectively, we define the *value* functions $\langle\langle 1 \rangle\rangle_{val}$ and $\langle\langle 2 \rangle\rangle_{val}$ for the players 1 and 2, respectively, as the following functions from the state space S to the set $\mathbb{R}$ of reals: for all states $s \in S$, let

$$\langle\langle 1 \rangle\rangle_{val}(f)(s) = \sup_{\sigma \in \Sigma} \inf_{\pi \in \Pi} \mathbb{E}_s^{\sigma,\pi}[f]; \quad \langle\langle 2 \rangle\rangle_{val}(-f)(s) = \sup_{\pi \in \Pi} \inf_{\sigma \in \Sigma} \mathbb{E}_s^{\sigma,\pi}[-f].$$

In other words, the value $\langle\langle 1 \rangle\rangle_{val}(f)(s)$ gives the maximal expectation with which player 1 can achieve her objective f from state s, and analogously for player 2. The strategies that achieve the value are called optimal: a strategy σ for player 1 is *optimal* from the state s for the objective f if $\langle\langle 1 \rangle\rangle_{val}(f)(s) = \inf_{\pi \in \Pi} \mathbb{E}_s^{\sigma,\pi}[f]$. The optimal strategies for player 2 are defined analogously. We now state the classical determinacy results for $2\frac{1}{2}$-player games with measurable objectives.

Theorem 1 (Quantitative determinacy [6]) *For all $2\frac{1}{2}$-player game graphs $G = ((S,E),(S_1,S_2,S_P),\delta)$ and for all measurable functions f, we have $\langle\langle 1 \rangle\rangle_{val}(f)(s) + \langle\langle 2 \rangle\rangle_{val}(-f)(s) = 0$ for all states $s \in S$.*

The determinacy result follows for two-level discounted games. In the following section we will study the complexity of optimal strategies and the computational complexity of solving two-level discounted games. We first recall a result about the classical discounted games. The classical discounted games are special cases of two-level discounted games such that $S_l = \emptyset$; i.e., the game consists of only upper level states. We refer to this class of games as *one-level* discounted games.

Theorem 2 (Memoryless determinacy of one-level discounted games [4]) *For all $2\frac{1}{2}$-player one-level discounted games, pure memoryless optimal strategies exist for both players.*

3 Strategy and Computational Complexity

We first show that pure memoryless optimal strategies exist in two-level discounted games. We first present a special class of two-level discounted games and reduce it to one-level discounted games.

One-step two-level discounted games. The class of one-step two-level discounted games are the special case of two-level discounted games such that the following restrictions are satisfied: (a) $S_l \subseteq S_P$ (i.e., every lower level state is a probabilistic state); and (b) $E \cap S_l \times S \subseteq S_l \times S_u$ (i.e., every successor of a state in S_l is a state in S_u). In other words, in one-step two-level discounted games from any lower level state the upper level states are reached in one step with probability 1. An one-step two-level discounted game can be reduced to one-level discounted games as follows. We convert every state in S_l to a state in S_u; and the reward function is modified as follows: we add rewards to the states in S_l to take care of the extra discounting step for converting a state in S_l to a state in S_u, i.e., for a state $s \in S_l$ its reward is assigned as

$$r(s) = (1-\beta) \cdot \sum_{t \in S} r(t) \cdot \delta(s)(t).$$

Since we can reduce one-step two-level discounted games to one-level discounted games, the existence of pure memoryless optimal strategies in one-step two-level discounted games follows.

Theorem 3 *Pure memoryless optimal strategies exist for both players in two-level discounted games.*

Proof. To prove the results we will use the existence of pure memoryless optimal strategies in reachability games, and present a reduction to one-step two-level discounted games.

First, for states in S_l we consider a reachability game as follows: once the game reaches a state $s \in S_u$, then player 1 receives the payoff $\langle\langle 1 \rangle\rangle_{val}(\text{TwoDisc})(s)$ and the game stops. The goal of player 1 is to maximize the payoff. Since this game is a reachability game, from the existence of pure memoryless optimal strategies in turn-based probabilistic reachability games [1], it follows that pure memoryless optimal strategies σ^* and π^* exist for both players in this game. Since the reward function is positive, it follows that $\langle\langle 1 \rangle\rangle_{val}(\text{TwoDisc})(s) > 0$ for all $s \in S_u$. Since in two-level discounted games player 1 can ensure to reach S_u with probability 1, it follows that once σ^* and π^* are fixed from all states in S_l states in S_u are reached with probability 1. Let T denote the random time when the game first reaches a state in S_u, and Θ_T denote the random variable for the T-th state. The strategies σ^* and π^* ensure the following:

1. for all strategies π and for all states $s \in S_l$ we have

$$\langle\langle 1 \rangle\rangle_{val}(\text{TwoDisc})(s) \geq \sum_{t \in S_u} \Pr_s^{\sigma^*,\pi}(\Theta_T = t) \cdot \langle\langle 1 \rangle\rangle_{val}(\text{TwoDisc}(t));$$

2. for all strategies σ and for all states $s \in S_l$ we have

$$\langle\langle 1 \rangle\rangle_{val}(\text{TwoDisc})(s) \leq \sum_{t \in S_u} \Pr_s^{\sigma,\pi^*}(\Theta_T = t) \cdot \langle\langle 1 \rangle\rangle_{val}(\text{TwoDisc}(t));$$

   and

3. $\Pr_s^{\sigma^*,\pi^*}(T < \infty) = 1$ and $\langle\langle 1 \rangle\rangle_{val}(\text{TwoDisc})(s) = \sum_{t \in S_u} \Pr_s^{\sigma^*,\pi^*}(\Theta_T = t) \cdot \langle\langle 1 \rangle\rangle_{val}(\text{TwoDisc}(t))$.

We now present a reduction from two-level discounted games to one-step two-level discounted games. We replace each state $s \in S_l$ as a probabilistic state such that from s the successor state distribution is as follows: $\delta(s)(t) = \Pr_s^{\sigma^*,\pi^*}(\text{Reach}(t))$ for $t \in S_u$. The following assertions hold in the one-step two-level discounted game for states in S_u:

1. for all states $s \in S_1 \cap S_u$, we have

$$\langle\langle 1\rangle\rangle_{val}(\text{TwoDisc})(s) = \max_{t \in E(s)} \beta \cdot r(s) + (1-\beta) \cdot \langle\langle 1\rangle\rangle_{val}(\text{TwoDisc}(t));$$

2. for all states $s \in S_2 \cap S_u$, we have

$$\langle\langle 1\rangle\rangle_{val}(\text{TwoDisc})(s) = \min_{t \in E(s)} \beta \cdot r(s) + (1-\beta) \cdot \langle\langle 1\rangle\rangle_{val}(\text{TwoDisc}(t));$$

and

3. for all states $s \in S_P \cap S_u$, we have

$$\langle\langle 1\rangle\rangle_{val}(\text{TwoDisc})(s) = \beta \cdot r(s) + (1-\beta) \cdot \sum_{t \in S} \delta(s)(t) \cdot \langle\langle 1\rangle\rangle_{val}(\text{TwoDisc}(t)).$$

The following assertions hold in the one-step two-level discounted game for states in S_l:

1. for all states $s \in S_1 \cap S_l$, we have $\langle\langle 1\rangle\rangle_{val}(\text{TwoDisc})(s) = \max_{t \in E(s)} \langle\langle 1\rangle\rangle_{val}(\text{TwoDisc}(t))$;
2. for all states $s \in S_2 \cap S_u$ we have $\langle\langle 1\rangle\rangle_{val}(\text{TwoDisc})(s) = \min_{t \in E(s)} \langle\langle 1\rangle\rangle_{val}(\text{TwoDisc}(t))$; and
3. for all states $s \in S_P \cap S_u$, we have $\langle\langle 1\rangle\rangle_{val}(\text{TwoDisc})(s) = \sum_{t \in S} \delta(s)(t) \cdot \langle\langle 1\rangle\rangle_{val}(\text{TwoDisc}(t))$.

From the above inequalities, the classical correctness proof for one-level discounted games, and the reduction of one-step two-level discounted games to one-level discounted games, it follows that the values in the original game and the reduced one-step two-level discounted games coincide. The combination of the pure memoryless optimal strategy in the discounted game obtained after reduction, and the pure memoryless optimal strategy in the reachability game is a witness optimal strategy in the two-level discounted game. Hence the result follows. ∎

Solution for two-level discounted MDPs. The existence of pure memoryless optimal strategies for two-level discounted games is proved by combining the solution of a reachability game and one-level discounted games. The result for MDPs follows as a special case. The value function for MDPs can be obtained from the solution of a linear programming problem which combines the linear programming solution for MDPs with reachability and one-level discounted objectives. The linear program for player-1 MDPs is as follows: the objective function is $\min_{s \in S} x_s$ subject to the following constraints

$$\begin{array}{lcll} x_s & \geq & x_t & s \in S_l \cap S_1; (s,t) \in E; \\ x_s & = & \sum_{t \in S} \delta(s)(t) \cdot x_t & s \in S_l \cap S_P; \\ x_s & \geq & \beta \cdot r(s) + (1-\beta) \cdot x_t & s \in S_u \cap S_1; (s,t) \in E; \\ x_s & = & \beta \cdot r(s) + (1-\beta) \cdot \sum_{t \in S} \delta(s)(t) \cdot x_t & s \in S_l \cap S_P; \end{array}$$

The solution for player-2 MDPs is similar. This gives us the following result.

Theorem 4 *Given a two-level discounted game on a player-1 MDP or a player-2 MDP, the value at all states can be computed in polynomial time.*

A class of discounted games has the *ordered field property* if for every game TwoDisc in the class with rewards, transition probabilities, and discount factors chosen from a field F, we have that the value $\langle\langle 1\rangle\rangle_{val}(\text{TwoDisc})(s)$ is also in F for each state s.

Corollary 1 (Ordered field property) *Given a two-level discounted game, if the rewards, discount factor, and transition probabilities are rational, then the value at evey state is rational. The class of all two-level discounted games have the ordered field property.*

Proof. The results follows from the existence of pure memoryless optimal strategies and the existence of linear program that characterizes the values on MDPs. Once a pure memoryless optimal strategy is fixed we have an MDP, and by the linear program characterizing the value for MDPs it follows that if the rewards, discount factor and transition probabilities are rational, then the value at every state is rational. The ordered field property follows from similar arguments. ∎

Complexity of two-level discounted games. Since pure memoryless optimal strategies exist for both players in two-level discounted games, and MDPs with two-level discounted objectives can be solved in polynomial time, it follows that the decision problem for the value function in two-level discounted games can be solved in NP ∩ coNP. Hence we have the following result.

Theorem 5 *Given a two-level discounted game, a rational number q, a state s, and $\bowtie \in \{\geq, >, \leq, <, =\}$, whether $\langle\langle 1\rangle\rangle_{val}(\text{TwoDisc})(s) \bowtie q$ can be decided in NP ∩ coNP.*

Algorithm for computing values. The existence of pure memoryless strategies ensure the correctness of the following naive algorithm to compute the values in two-level discounted game: (a) enumerate all pure memoryless strategies, and for each pure memoryless strategy compute the value for the MDP obtained by fixing the strategy (using the linear program), and (b) choose the value of the best pure memoryless strategy. The above algorithm is an exhaustive search on the set of pure memoryless strategies. We now describe an efficient search on the set of pure memoryless strategies given as a strategy improvement algorithm for two-level discounted games. The strategy improvement algorithm combines in a hierarchical fashion two classical strategy improvement algorithms: (a) the strategy improvement algorithm for stochastic games with discounted objectives [4] and (b) the strategy improvement algorithm for stochastic reachability games [2].

The strategy improvement algorithm is as follows: (a) fix a pure memoryless strategy at the upper-level states; (b) apply the strategy improvement algorithm for reachability games for the lower-level reachability game to compute values given the strategy that is fixed in the higher-level game; and (c) once the values are computed, apply the strategy improvement step for discounted games to improve the upper-level strategy. The algorithm stops when no improvement is possible and obtains a pure memoryless optimal strategy. This gives us a strategy improvement algorithm to compute values in two-level discounted games.

4 Conclusion

We have introduced a new model of stochastic games that provide a uniform framework for decision making across different time scales. We have shown that pure memoryless optimal strategies exists in these games. Our framework subsumes classical discounted games, and provides a natural extension in which discounting is applied at different time granularities. We show that in our framework the solution for MDPs can be achieved in polynomial time matching the best known bound of MDPs with discounted objectives. For two-level turn-based stochastic games we show that whether the value is equal to a rational can be decided in NP ∩ coNP, matching the best known complexity bound for discounted stochastic games.

References

[1] A. Condon (1992): *The Complexity of Stochastic Games*. *Information and Computation* 96(2), pp. 203–224.

[2] A. Condon (1993): *On Algorithms for Simple Stochastic Games*. In: *Advances in Computational Complexity Theory, DIMACS Series in Discrete Mathematics and Theoretical Computer Science* 13, American Mathematical Society, pp. 51–73.

[3] S. Dziembowski, M. Jurdzinski & I. Walukiewicz (1997): *How much memory is needed to win infinite games?* In: *LICS'97*, IEEE, pp. 99–110.

[4] J. Filar & K. Vrieze (1997): *Competitive Markov Decision Processes*. Springer-Verlag.

[5] P.R. Kumar & T.H. Shiau (1981): *Existence of Value and Randomized Strategies in Zero-sum Discrete-Time Stochastic Dynamic Games*. *SIAM J. Control and Optimization* 19(5), pp. 617–634.

[6] D.A. Martin (1998): *The determinacy of Blackwell games*. *The Journal of Symbolic Logic* 63(4), pp. 1565–1581.

[7] L.S. Shapley (1953): *Stochastic Games*. *Proc. Nat. Acad. Sci. USA* 39, pp. 1095–1100.