Making the Best of Limited Memory in Multi-Player Discounted Sum Games

Anshul Gupta and Sven Schewe
anshul@liverpool.ac.uk, sven.schewe@liverpool.ac.uk

Department of Computer Science
University of Liverpool, Liverpool, UK

Abstract. In this paper, we establish the existence of optimal bounded memory strategy profiles in multi-player discounted sum games. We introduce a non-deterministic approach to compute optimal strategy profiles with bounded memory. Our approach can be used to obtain optimal rewards in a setting, where a powerful player selects the strategies of all players for Nash and leader equilibria, where in leader equilibria the Nash condition is waived for the strategy of this powerful player. The resulting strategy profiles are optimal for this player among all strategy profiles that respect the given memory bound, and the related decision problem is NP-complete. We also provide simple examples, which show that having more memory will improve the optimal strategy profile, and that sufficient memory to obtain optimal strategy profiles cannot be inferred from the structure of the game.

1 Introduction

Discounted sum games [14][13] are the stochastic games with quantitative objectives that have been introduced by Shapley [14]. They are played on a finite directed graph without sinks, where each vertex is owned by one of the players. Intuitively, they are played by placing a token on the graph, which is moved forward by the players. Initially, the token is placed on the initial vertex. Whenever the token is on a vertex, the player who owns this vertex will select an outgoing edge and move the token along this edge. This way, the players construct an infinite play. Quantitative games [3] are good models to study non-terminating programs with multiple components that interact in non-cooperation mode. In quantitative games, players have goals defined by the payoffs on the edges (sometimes on the vertices). For these payoffs, the players have quantitative targets, such as maximising their individual limit average or the discounted sum of their individual rewards, where the value of a play is computed under a discount factor.

Solutions to these games are the strategy profiles that consists of strategies—recipes how to play—for each player. A realistic situation however, needs to be implementable, and thus have to cope with the limited resources such as limited memory. Strategy profiles should also satisfy the consistency constraints. The reason for the latter is that players are assumed to be rational. The lowest level of rationality is to take a look at the strategy profile and check if one would gain by changing the own strategy. Strategy profiles where all strategies pass this test are stable in terms of Nash equilibria [10][11][12].
That is, a strategy profile is in a Nash equilibrium, if no player benefits from changing her strategy unilaterally.

We have recently introduced a more general and broader class of strategy profiles than Nash equilibria, called leader strategy profiles \[9\], in the context of multi-player mean pay-off games \[16,4\]. In leader strategy profiles, we assume that a powerful player may select a strategy profile that satisfies Nash’s stability requirement only for the other players. This player, the leader or dictator for short, can assign the strategies to all players, including herself. While we still require the strategy profile to be stable in that the other players do not have an incentive to deviate, the dictator herself may be in a position to improve over her current strategy by deviating unilaterally. Thus, all Nash equilibria are leader strategy profiles, but not all leader strategy profiles are Nash. We call leader strategy profiles that are optimal for the leader leader equilibria. The more relaxed condition of a leader strategy profile implies that leader equilibria can be selected from a larger base, and are also superior to Nash equilibria as we show in section \[8\]. In this paper, we study leader equilibria and Nash equilibria for the leader in discounted sum games that uses bounded memory.

### 1.1 Results

This paper extends the use of leader equilibria to multi-player discounted sum games. We focus on the strategy profiles with bounded memory. We show that leader equilibria always exist, even when the strategies are restricted to memoryless ones. We also show that memory can help improving the strategy, and that no bound on the memory needed for optimal strategy profiles can be inferred from the structure of the game graph.

We show that considering reward and punish strategy profiles \[7,3,9\] is sufficient. These are the strategy profiles where, as soon as a player deviates from the strategy assigned to her by the dictator, all other players henceforth team up against this player, prioritising the goal to harm the deviated player over increasing their own payoff by never entering the part of the game when all players follow their assigned strategy: this behaviour has no negative effect on their payoff, but merely strengthens the means of the dictator to keep the individual players in line. We introduce a simple non-deterministic polynomial time approach for assigning strategies that meets or exceed a given payoff bound for the dictator and uses only memory within a given bound. We note that the decision problem whether a pure positional strategy with bounded memory that gives a reward greater than or equal to some threshold value exists is NP-complete.

### 1.2 Related Work

The theory of stochastic games was introduced by Shapley in \[14\]. He shows that for the two players, discounted sum game has a value and the optimal positional strategies exist for both the players. This idea is further extended in \[5\] to give the stationary equilibrium in stochastic multi-player game. Bewley and Kohlberg \[2\] show that in a two-player infinite stochastic game, where both action and state spaces are finite, stationary optimal strategies exist for both the players. Gimbert and Zielonka \[8\] studied infinite two-player antagonistic games and give conditions that assure for both players
from initial sequences of plays to actions, from a bounded initial value \( \sigma \). For a bounded memory strategy, the choice of the next vertex depends only on the current position, whereas a bounded-positional strategy is a strategy, in which the choice of the next vertex depends on finite memory. A positional strategy is a strategy, in which the respective player chooses the successor vertex defined by a \( \pi \) from initial sequences of plays that end in some vertex of the graph. We consider two types of strategies, positional and bounded memory strategies. A positional strategy is a strategy, in which the choice of the next vertex depends only on the current position, whereas a bounded-memory strategy is a strategy, where the choice of next vertex depends on finite memory. For a bounded memory \( M \) (where \( M \) is simply a finite set of fixed size, the memory bound with a dedicated initial value \( m \)), we define the memory update function \( \mathcal{M} \) as \( \mathcal{M} : M \times (V \times A) \rightarrow M \). Thus, our memory works as a Moore machine without output.

2 Preliminaries

A multi-player discounted sum game (MDSG) is a game played on the finite directed weighted graph \( G \) defined as a tuple \( (P, V, \{V_p \mid p \in P\}, v_0, A, T, \{r_p : V \times A \rightarrow \mathbb{Q} \mid p \in P\}) \), where \( P \) is a set of players, \( V \) is a set of vertices with a designated initial Vertex \( v_0 \in V \), \( \{V_p \mid p \in P\} \) is a partition of the vertices \( V \) into the sets \( V_p \) of vertices owned by player \( p \), \( A \) is a finite set of actions, \( T : V \times A \rightarrow V \) is a set of transitions that maps vertices and actions to vertices, and \( \{r_p \mid p \in P\} \) is a family of reward functions defined as \( r_p : V \times A \rightarrow \mathbb{Q} \) for all \( p \in P \) that assigns, for each respective player \( p \), a reward for each action \( a \) that is taken from a Vertex \( v \) (or, likewise, for the transition taken). The game is played by moving a token along the edges of the graph, starting from the initial Vertex \( v_0 \). Each Vertex \( v \) belongs to exactly one player \( p \). At Vertex \( v \), the player who owns \( v \) selects the next action \( a \). The pebble is then moved forward to the vertex as given by the transition \( T(v, a) \). We denote the reward for player \( p \) at any transition \( T(v, a) \) by \( r_p(v, a) \). An MDSG is called a zero-sum game if, for all vertices \( v \in V \) and for all actions \( a \in A \), \( \sum_{p \in P} r_p(v, a) = 0 \) holds. This results in an infinite path, called a play. The payoff at every transition is discounted by a discount factor \( \lambda \), where \( 0 < \lambda < 1 \). In discounted sum games, the payoff (or: reward) for player \( p \) at the \( i^{th} \) transition is given by \( r_p(v_i, a_i) \cdot \lambda^i \). Thus, for an infinite play \( \pi = v_0, a_0, v_1, \ldots \), rewards for player \( p \) is evaluated to

\[
r_p(\pi) = \sum_{i=0}^{\infty} r_p(v_i, a_i) \cdot \lambda^i
\]

The way that the respective player \( p \) chooses the successor vertex is defined by a strategy \( \sigma_p \). We distinguish pure strategies, which are functions \( \sigma_p : (V A)^* \rightarrow A \) from initial sequences of plays to actions, from mixed strategies, which are functions \( \sigma_p : (V A)^* \rightarrow \text{dist}(A) \) from initial sequences of plays that end in some vertex of player \( p \) to a distribution over the actions in \( A \). We consider two types of strategies, positional and bounded memory strategies. A positional strategy is a strategy, in which the choice of the next vertex depends only on the current position, whereas a bounded-memory strategy is a strategy, where the choice of next vertex depends on finite memory. For a bounded memory \( M \) (where \( M \) is simply a finite set of fixed size, the memory bound with a dedicated initial value \( m \)), we define the memory update function \( \mathcal{M} \) as

\[
\mathcal{M} : M \times (V \times A) \rightarrow M
\]

Thus, our memory works as a Moore machine without output.
where $M$ is the memory and $\mathcal{M}$ is the transition function. The input alphabet $V \times A$ is a product of the last vertex, the action selected, and the vertex reached on a transition. A family of strategies $\sigma = \{\sigma_p \mid p \in P\}$ is called a strategy profile. A strategy profile $\sigma$ defines an expected reward, denoted $\mathbb{E}_p(\sigma)$ for each player $p$. In this paper, we shall focus on the reward of positional and bounded memory strategy profiles. For a positional strategy profile $\sigma$, the payoff from every vertex is well defined. By abuse of notation, we use $\mathbb{E}_p(\sigma, v) = \sum_{a \in A} \sigma(v)(a) \cdot \left( r_p(v, a) + \lambda \mathbb{E}_p(\sigma, T(v, a)) \right)$, to denote the payoff for player $p$ when starting in a vertex $v$. Note that this implies $\mathbb{E}_p(\sigma) = \mathbb{E}_p(\sigma, v_0)$.

A strategy profile is a Nash equilibrium if no player has an incentive to change her strategy, provided that all other player keep theirs. That is, for all players $p \in P$ and for all $\sigma' = \{\sigma'_q \mid q \in P\}$ with $\sigma_q = \sigma'_q$ for all $q \neq p$, $\mathbb{E}_p(\sigma) \geq \mathbb{E}_p(\sigma')$ holds. A strategy profile is a leader strategy profile [9] for a designated player $d$ (for dictator), if no other player has an incentive to deviate her strategy. That is, if, for all players $p \in P \setminus \{d\}$ and for all $\sigma' = \{\sigma'_q \mid q \in P\}$ with $\sigma_q = \sigma'_q$ for all $q \neq p$, $\mathbb{E}_p(\sigma) \geq \mathbb{E}_p(\sigma')$ holds. A Nash resp. leader strategy profile is optimal for a class of strategies and is a Nash resp. leader equilibrium, if no other strategy profile of this class gives a higher payoff for the dictator. In two-player discounted sum games (DSGs), the set of vertices in $\mathcal{G}$ is partitioned into two sets where each vertex belongs to exactly one of the players and the player who owns the vertex decides the next move. For a multi-player DSG $\mathcal{G} = \langle P, V, \{V_p \mid p \in P\}, v_0, A, T, \{r_p : V \times A \to \mathbb{Q} \mid p \in P\} \rangle$, we define the two-player zero-sum DSG $\mathcal{G} = \langle P, V, \{V_p, V_o\}, v_0, A, T, \{r_p, r_o\} \rangle$ played between $p$ and an opponent $o$, where the nodes of $p$ and $o$ partition $V$ into two sets ($V_o = V \setminus V_p$) and their goals are antagonistic ($r_o(v, a) \mapsto -r_p(v, a)$). (Note that not all multi-player DSGs with two players in game are two-player games in this sense. Two player games need to be antagonistic zero-sum games). A game is called memoryless determined if all players have optimal memoryless strategies. Two-player discounted sum games are memoryless determined [16]: both players have an optimal positional strategy.

**Theorem 1.** [16] Two-player discounted sum games are memoryless determined.

We denote the expected outcome for player $p$ in a two-player game that starts at any vertex $v$ by $r_p(v)$.

### 3 Leader and Nash equilibria

In this section, we show that leader equilibria are superior to Nash equilibria in simple zero-sum discounted sum games. To show this, consider the three-player game from Figure 1. One of the players, Player 2, acts as the dictator. The game is played on a simple graph with three vertices, named 1, 2, and 3, owned by the respective player with the same name. The game graph with the payoff vectors of each transition is shown in Figure 1 and we use a discount factor of $\lambda = \frac{1}{3}$. The payoff vectors represent the payoff of Player 1, the dictator, and Player 3, in this order. Initially, Player 1 can choose to play to Vertex 2 or she can choose to remain in Vertex 1. She plays to Vertex 2 only if the dictator, in her strategy profile, chooses to remain in Vertex 2 for a while.
At Vertex 2, the dictator has different options. She can choose to play to Vertex 3 (this is the option where she maximises her reward), she can choose to remain in 2 for a while, before continuing to Vertex 3, or she can stay in Vertex 2 forever. It is easy to notice that, when in Vertex 2, the dictator will immediately continue to Vertex 3 in all Nash equilibria. Consequently, Player 1 would never play to Vertex 2 from Vertex 1: staying in Vertex 1 for ever will yield a payoff of 0, while moving to Vertex 2 in round $i$ would, for $\lambda = \frac{1}{2}$, result in a payoff of $-3^i$. Thus, the only play that can result from a Nash equilibrium is the play $1^\omega$, where the overall reward for all participating players is 0. A leader equilibrium, however, is, for the dictator to stay twice in Vertex 2 and then progress to Vertex 3. In this case, the dictator can assign Player 1 the strategy to immediately progress to Vertex 2, resulting in the play $1^1, 2^2, 2^2, 3^\omega$. This will provide an overall payoff of 0 for Player 1, 1.5 for the dictator, and $-1.5$ for Player 3.

**Theorem 2.** Compared to Nash equilibria, leader equilibria may result in higher, but will never provide smaller rewards for the dictator.

While the example has proven the ‘higher’ part, note for the ‘not smaller’ part that all Nash equilibria are leader equilibria, such that a leader equilibrium cannot be inferior to a Nash equilibrium. They can, of course, be equal when a leader equilibrium is Nash. This is, for example, the case when dictator owns no vertex. Note that the game from Figure 1 can be used to argue that having memory helps, and having more memory helps more. Among the positional strategies of the dictator, staying in Vertex 2 forever (with an overall payoff of 1 for Player 1 and the dictator, and $-2$ for Player 3, respectively) is superior to continuing immediately to Vertex 3 (because in the latter case Player 1 will stay in Vertex 1, see above). So, while still superior to the only Nash equilibrium, it is inferior to the strategy described above, which uses a tiny amount of memory. To observe that, in general, more memory helps more, consider the situation where one lets $\lambda$ grow towards one. It is easy to see that, the closer $\lambda$ gets to one, the longer dictator would stay in Vertex 2 in leader equilibrium for the respective discount factor. The optimal memory bounded strategy for the dictator therefore improves with the memory we allow for.

**Lemma 1.** The quality of Nash and leader equilibria improves with the size of the available memory.

It now becomes tempting to assume that we could use this observation to identify a situation where an optimal leader strategy profile is reached, i.e., Given a game graph and a discount factor, is there a $k \in \mathbb{N}$ such that an optimal leader strategy profile for memory $k$ is considered optimal for infinite memory? The answer to this question is negative. In the example from Figure 1, we argue that having a finite memory at the vertices is sufficient for a leader equilibrium, but the effect of increasing the memory is different than in our first example. Irrespective of the discount factor it is apparent that the dictator needs to promise sufficiently many, say $s$, loops in Vertex 2 so that
\[ \sum_{i=0}^{s-1} \lambda^i \geq \frac{1-\varepsilon}{1-\lambda}. \]

Consequently, the number of repetitions grows to infinity, for all \( \lambda \in [0, 1] \), and with \( \varepsilon \) falling to 0.

If the memory is smaller than minimal such \( s \), then the dictator would receive an overall reward of 0, either because she promises to stay for more than the memory bound many steps (and thus for ever) in Vertex 2, or by not promising to do so and hence tempting the first player to move to Vertex 4. If, on the other hand, the memory size is at least \( s \), then the dictator has enough memory to play the optimal pure strategy to move to Vertex 3 after \( s \) loops in Vertex 2.

**Theorem 3.** For a given size and a discount factor \( \lambda \), there is no memory bound \( k \) such that an optimal leader strategy profile with memory bound \( k \) is an optimal leader strategy profile.

### 4 Reward and punish strategy profiles in discounted sum games

In this section, we show that for a play \( \pi \), we could establish if there exists a leader (or Nash) strategy profile \( \sigma \) with \( \pi = \pi_\sigma \), and that, moreover, its extension to such a strategy profile is simple. For this, we first introduce reward and punish strategy profiles.

In reward and punish strategy profiles [9], the dictator assigns a strategy to each player and each of them co-operates to produce a play \( \pi \) while playing in accordance with the assigned strategies. As soon as one player deviates, the remaining players team up with dictator and co-operate against the deviating player \( i \). That is, they will henceforth follow the goal to minimise the payoff of player \( i \), and act jointly as the antagonist of \( i \) in the underlying two-player DSG. Thus, in the resultant two-player game, while the objective of player \( i \) is still the same, the objective of all other players (including dictator), is changed and has become to minimise the pay-off of player \( i \). Assuming that positional optimal strategies in this two-player DSG are fixed, \( \pi \) thus defines a reward and punish strategy profile, which we denote by \( rps(\pi) \).

We now argue that

1. every leader resp. Nash strategy profile \( \sigma \) can be transformed into a leader resp. Nash strategy profile \( \sigma' \) with \( \pi_\sigma = \pi_{\sigma'} \), and thus with similar rewards for all players, and
2. give necessary and sufficient conditions for a play \( \pi \) to be defined by some leader resp. Nash strategy profile.

We first discuss the necessary conditions for a path to be the outcome of a Nash resp. leader equilibrium, and then show that it is sufficient for a path to be the outcome of a Nash resp. leader reward and punish strategy profile.

**Lemma 2.** If \( \pi = v_0, a_0, v_1, \ldots \) is the outcome of a Nash resp. leader equilibrium, then, for all \( j \in \mathbb{N} \) and all players \( p \) resp. all players \( p \neq d \),

\[ r_p(v_j) \leq \sum_{i=0}^{\infty} r_p(v_{j+i}, a_{j+i}) \cdot \lambda^i \] holds.
Proof. We assume for contradiction that the condition is violated. We therefore select a \( j \in \mathbb{N} \), and a player \( p \) (for leader equilibria a player \( p \neq d \) ) such that \( r_p(v_j) > \sum_{i=0}^{\infty} r_p(v_{j+i}, a_{j+i}) \cdot \lambda^i \). We then change the strategy of player \( p \) to follow her strategy from the two-player DSG from position \( j \) onwards. The resulting play \( \pi' = v_0', a_0', v_1', \ldots \) with \( v_i' = v_i \) for all \( i \leq j \) and \( a_i' = a_i \) for all \( i < j \) satisfies

\[
\begin{align*}
r_p(\pi') &= \sum_{i=0}^{\infty} r_p(v_i', a_i') \cdot \lambda^i \\
&= \sum_{i=0}^{j-1} r_p(v_i', a_i') \cdot \lambda^i + \lambda^j \sum_{i=0}^{\infty} r_p(v_{j+i}, a_{j+i}) \cdot \lambda^i \\
&\geq \sum_{i=0}^{j-1} r_p(v_i, a_i) \cdot \lambda^i + \lambda^j r_p(v_j) \\
&> \sum_{i=0}^{\infty} r_p(v_i, a_i) \cdot \lambda^i + \lambda^j \sum_{i=0}^{\infty} r_p(v_{j+i}, a_{j+i}) \cdot \lambda^i \\
&= \sum_{i=0}^{\infty} r_p(v_i, a_i) \cdot \lambda^i = r_p(\pi).
\end{align*}
\]

Lemma 3. If \( \pi = v_0, a_0, v_1, \ldots \) satisfies \( r_p(v_j) \leq \sum_{i=0}^{\infty} r_p(v_{j+i}, a_{j+i}) \cdot \lambda^i \) for all \( j \in \mathbb{N} \) and all players \( p \) resp. all players \( p \neq d \), then \( rps(\pi) \) is a Nash resp. leader equilibrium.

Proof. We assume for contradiction that a player \( p \) (for leader equilibria a player \( p \neq d \) ) has an incentive to deviate, and that the first position where player \( p \) selects a different action is \( j \in \mathbb{N} \). Let \( \pi' = v_0', a_0', v_1', \ldots \), where \( v_i' = v_i \) for all \( i \leq j \) and \( a_i' = a_i \) for all \( i < j \), be the resulting play. We have

\[
\begin{align*}
r_p(\pi') &= \sum_{i=0}^{\infty} r_p(v_i', a_i') \cdot \lambda^i \\
&= \sum_{i=0}^{j-1} r_p(v_i', a_i') \cdot \lambda^i + \lambda^j \sum_{i=0}^{\infty} r_p(v_{j+i}, a_{j+i}) \cdot \lambda^i \\
&\leq \sum_{i=0}^{j-1} r_p(v_i, a_i) \cdot \lambda^i + \lambda^j r_p(v_j) \\
&\leq \sum_{i=0}^{\infty} r_p(v_i, a_i) \cdot \lambda^i + \lambda^j \sum_{i=0}^{\infty} r_p(v_{j+i}, a_{j+i}) \cdot \lambda^i \\
&= \sum_{i=0}^{\infty} r_p(v_i, a_i) \cdot \lambda^i = r_p(\pi).
\end{align*}
\]

The first ‘\( \leq \)’ is implied by the definition of \( rps \), as the remaining players will play antagonistic to \( p \), such that \( p \) cannot yield a better result than \( r_p(v_j) \) starting from \( v_j \).

Together with the observation that pure Nash equilibria always exist [3]—leader equilibria can be formed by all players (playing as if they played their respective two-player DSG)—these lemmas provide the following theorem.

Theorem 4. Pure Nash and leader strategy profiles always exist in MDSGs, and for finding optimal ones, it suffices to consider reward and punish strategies.

This is particularly interesting when we focus on the implementable strategy profiles. In order for a strategy to be implementable, it needs to be realizable with finite memory, and we are particularly interested in finite memory strategies with a given small bound \( b \) on the memory used. Note that, for reward and punish strategy profiles, we do not have to record the reaction upon deviation, as it is implicitly described by the punishment part. Thus, we do not want to reason about the trivial part in the strategy, and therefore do not count the tiny bit of memory required for the punishment part. Note that this part does not need much memory: it suffices to remember which player is responsible for the deviation and at which vertex. When we allow for finite memory \( M \), this effectively defines a larger game, on which a memoryless strategy is used. For
a game \( G = \langle P, V, \{ V_p \mid p \in P \}, v_0, A, T, \{ r_p : V \times A \rightarrow Q \mid p \in P \} \rangle \) and finite memory \( M \) with initial memory \( m_0 \in M \), we can simply define \( G^M = \langle P, V', \{ V'_p \mid p \in P \}, v'_0, A, T, \{ r'_p : V' \times A \rightarrow Q \mid p \in P \} \rangle \) with \( V' = V \times M \), \( V'_p = V_p \times M \), \( v'_0 = (v_0, m_0) \), and \( r'_p : ((v, m), a) \mapsto r_p(v, a) \).

**Corollary 1.** Pure positional, and, for a given memory bound \( b \), bounded memory Nash and leader strategy profiles always exist in MDSGs, and for finding the optimal ones, it suffices to consider reward and punish strategies.

## 5 Constraints for finite pure reward and punish strategy profiles

We first state that optimal strategies exist for all memory bounds. This is a simple implication of Theorem 4 and the finite space of candidate strategy profiles.

**Lemma 4.** For all MDSGs and for all memory bounds, optimal strategy profiles exist among the Nash and leader equilibria.

We infer a necessary and sufficient constraint system for the strategy profiles in Nash and leader equilibria in MDSGs. Theorem 4 implies that, whenever a player deviates at some vertex \( v \), then the remainder of the game resembles a two-player game that starts at \( v \). The player who owns vertex \( v \) therefore has an incentive to deviate if, and only if, her payoff from now onwards would be less than the payoff she receives in this underlying two-player game. This provides us with a first necessary constraint, namely

\[
- \text{ at any history } h \text{ that ends in a vertex } v, \text{ which is owned by player } p \in P, E_p(\sigma, h) \geq r_p(v).
\]

For positional reward and punish strategies \( \sigma \), the subtrees in all histories \( h \) that end in \( v \) coincide, such that one can write \( E_p(\sigma, v) \) instead of \( E_p(\sigma, h) \).

For pure strategies, we require for every vertex \( v \) that

\[
- \text{ there is an action } a \text{ such that, for all players } p \in P, E_p(\sigma, v) = \sum_{a \in A} \sigma(v)(a) \cdot \left( r_p(v, a) + \lambda E_p(\sigma, T(v, a)) \right),
\]

The action \( a \) from these constraints refers to the action selected at vertex \( v \) by player \( p \) in strategy profile \( \sigma \). Once these actions are fixed, we therefore have a simple linear equation system of full degree, that can easily be solved. To determine if the resulting system is in equilibrium we can simply check if the first set of constraints hold for all players (Nash equilibrium) or for all players but the dictator (leader equilibrium). To validate that there is a pure strategy profile of a predefined quality can therefore be checked in nondeterministic polynomial time.

**Lemma 5.** We can check if there is a positional strategy profile that meets or exceeds a given threshold \( t \) for the dictator reward and is a leader or Nash equilibrium in nondeterministic polynomial time.

\(^1\) Recall that we do not count in the little memory needed for the punishment case. Positional therefore only refers to the play \( \pi_e \).
For strategy profiles with bounded memory, we can simply use the extended memory game instead.

**Corollary 2.** We can check if there is a pure bounded memory strategy profile with fixed memory bound $b$ that meets or exceeds a given threshold $t$ for the dictator reward and is a leader or Nash equilibrium in nondeterministic polynomial time in the size of the extended memory game.

In order to establish NP completeness, we reduce the satisfiability of a 3SAT formula over $n$ atomic propositions with $m$ conjunctions to solving a multi-player DPG with $2n + 1$ players and $4m + 5n + 2$ vertices that uses only payoffs 0 and 1. This is a standard reduction, which is similar to the reduction for mean payoff games [9], safe for the weights.

We consider the reduction for the example of the 3SAT formula

$$ (p \lor \neg q \lor \neg r) \land \neg p \lor q \lor \neg r \land \neg p \lor q \lor r $$

The $2n + 1$ players consists of $2n$ players for the $2n$ literals corresponding to the $n$ variables, and the dictator who intuitively tries to validate the formula. The game consists of three phases, an initial assignment phase, in which the dictator intuitively assigns either the value true or the value false to all $n$ variables. We use two players for each of the variables, one representing true, and one representing false. In a second validation phase, the dictator intuitively validates that the chosen assignment indeed satisfies the specification $\varphi$. For this, she successively steps through the conjuncts of the 3SAT formula. For each conjunct, the dictator can select one of the three literals, which is owned by the same player who owns this literal in the first phase. In the first and second phase, the literal players can either continue, or move to an absorbing state. In a final evaluation phase, the game goes round a ring of length $n$, where, in each step, a disadvantage is given to one of the players of a variable, the player who represents true, or the player who represents false. By choosing the payoff for cycling in the absorbing state accordingly, we can assure that there is a leader / Nash equilibrium with payoff $> 0$ for the dictator if $\varphi$ is satisfiable, and a payoff of 0 for the dictator if $\varphi$ is unsatisfiable, using only the rewards 0 and 1.

**Assignment phase** In the assignment phase, we have two types of vertices. We have vertices 0 through $n$, which are owned by the dictator and 2$n$ literal vertices. In Vertex $i - 1$, the dictator chooses either the truth value or false value of each variable $Z_i$ by either moving to a Vertex $z_i$, or moving to $\neg z_i$, owned by the players with the respective literal. From Vertex $z_i$ or $\neg z_i$, the respective player can choose to move to the dictator vertex, or to an absorbing Vertex abs. From the absorbing Vertex abs, there is just one outgoing transition, which returns to abs, and has a payoff of 0 for dictator and payoff of 1 for all other players.

![Assignment phase](image.png)

**Fig. 3.** Assignment phase: All actions have a payoff of 0 for the dictator and 1 for every other player.
All other transitions in this phase have the same payoff: 0 for the dictator and 1 for every other player. The assignment phase is shown in Figure 3.

**Validation phase.** In the validation phase, the dictator intuitively tries to validate that her chosen assignment indeed validates the formula \( \varphi \). Here, we have two types of vertices. We have \( m \) dictator vertices and \( 3m \) literal vertices, \( z_1^i, z_2^i, \) and \( z_3^i \) for all \( i \in \{1, \ldots, m\} \). In this phase, the dictator successively steps through the conjuncts of the 3SAT formula. For each conjunct, the dictator selects one of the three literals, which is owned by the same player who owns this literal in the first phase. At Vertex \( n+i−1 \), the dictator can play to any literal vertex, to \( z_1^i \), to \( z_2^i \), or to \( z_3^i \) in conjunct \( i \), to validate the value of the conjunct.

Further, we have the same absorbing vertex as in the assignment phase. Here also, payoff at the edges that can be taken only once are omitted. The Validation phase is shown in the Figure 4.

**Evaluation phase.** In the evaluation phase, we have \( 2n \) literal vertices and a single dictator vertex. The evaluation phase of the multi-player DSG resembles a ring structure. Here, the game cycles in a ring of length \( n \) where at every vertex one of the players is at disadvantage. At any Vertex \( z_i \) resp. \( \neg z_i \), its counter-literal receives a payoff of 0 while everyone else receives a payoff of 1. For \( i \neq n \), the Vertex owned by \( z_i \) has two successors, the vertices owned by \( z_{i+1} \) and \( \neg z_{i+1} \). \( z_n \) has the dictator vertex as its only successor. The Vertex owned by \( \neg z_i \) has the same successor/s as the Vertex owned by \( z_i \): the vertices owned by \( z_{i+1} \) and \( \neg z_{i+1} \) for \( i < n \), and the dictator Vertex for \( \neg z_n \). The evaluation phase is shown in the Figure 5. If a satisfying assignment exists, the dictator intuitively selects one. The players that represent literals that agree with this assignment will receive a reward of \( 1 \), a reward \( < \frac{1}{1-\lambda} \) for the other literals, and the dictator a
reward > 0. If no such assignment exists, the dictator cannot reach the evaluation phase, as she would have to reduce the award of \( p \) or \( \neg p \) for each atomic proposition \( p \). But for some atomic proposition \( p \), both \( p \) and \( \neg p \) need to be passed by in the assignment phase, and one of the players has an incentive to deviate to the absorbing state abs. In this case, all literal players receive a payoff of \( \frac{1}{1-\lambda} \), while the dictator receives a payoff of 0.

**Lemma 6.** To check if there is a pure positional or bounded memory strategy profile with fixed memory bound \( b \) that meets or exceeds a given threshold \( t \) for the dictator and is a leader or Nash equilibrium is NP hard.

Together with Lemma 5 and Corollary 2, the NP hardness provides:

**Theorem 5.** To check if there is a pure positional or bounded memory strategy profile with fixed memory bound \( b \) that meets or exceeds a given threshold \( t \) for the dictator and is a leader or Nash equilibrium is NP complete.

### 6 Equilibria with extended observations

We first argue why we focus on pure strategies, despite the fact that mixed strategies are a more general choice. In principle, all arguments from the previous sections also extend to the randomised strategies and strategy profiles, such that one might argue to use the randomised model. The reason why we refrained from doing so is that reward and punish strategies are implementable: the punishment follows an observed deviation. Such temporal dependencies are *not* common in the definition of Nash equilibria. This is unsurprising when given their origin in concurrent games, where only a single move is played and the concept of history and temporal order of cause and effect does not apply. For us, the concept of implementability outweighs the generality of randomised strategies. The issue with randomisation is that the compliance with the randomised decision is not easily distinguishable from the decisions that only use similar or smaller supports; as long as each action would have been possible under randomised strategies, no violation would be observed. In this respect, an assigned strategy becomes non-deterministic rather than randomised. Each player can secretly change her decision, e.g., by making it deterministic. In all cases where any pair of possible outcomes gives different payoffs to the player who owns the current vertex, this would lead to a deviation. Assuming that the players are selfish, players might also be tempted to play deterministically if not observed, in order to save the trouble of producing a randomised decision.

The above argument driven by the unobservability of deviation in mixed strategies thus allows us to focus on pure strategies. However, an alternative to this restriction is to lift the restriction of our observational power: instead of observing the outcome of a decision, we observe the decision itself. Note that this would imply an uncountable set of possible actions, as encoded in the different selected probability distribution over actions, which are possible in every vertex. To justify making this observable, one might think of externalising how to resolve the probabilities, say, by a highly trusted third party. Also, note that allowing for mixed strategies does not remove the usefulness of memory. In the example from Figure 6, when in Vertex 1, the dictator can only assign an
equilibrium strategy to Player 1 which is not worse for Player 1 than staying in Vertex 1. Initially (that is, on the empty history), however, she does not have to take the interest of Player 1 into account and can progress to Vertex 1 with probability $1 \cdot \frac{1}{2}$.

Corollary 1 establishes that it suffices to focus on reward and punish strategy profiles. This implies a simple constraint system for extended memory games: no player (except for the dictator in leader equilibria) may reach a position, where a player would benefit from changing her strategy in a reward and punish strategy profile (that is assigned by the dictator). Thus, at every Vertex $v$ of the extended memory game with memory $m$, it must hold that $E_p(v, m) \geq E_{p,2}(v)$. Here, $E_p(v, m)$ is the expected reward for player $p$ at vertex $v$ in extended memory game and $E_{p,2}(v)$ is the expected reward for player $p$ at vertex $v$ in two-player game that would result if player $p$ chooses to deviate at vertex $v$. Apart from this, the normal equality constraints must hold, that is the payoff at each extended memory vertex is the expected payoff from the next transition plus the discounted expected payoff of the target positions, i.e.,

$$E_p(\sigma, v) = \sum_{a \in A} \sigma(v)(a) \cdot \left( r_p(v, a) + \lambda \sum_{v' \in V} T(v, a)(v') \cdot E_p(\sigma, v') \right).$$

We can again use a non-deterministic approach to solve the related decision problem. E.g., We can start by guessing a probability distribution at each vertex $v_i$ on all its outgoing actions, and guess, for each action, a target memory value. Once these distributions are fixed, we can again solve the resulting linear equation system, and simply check that it satisfies the constraints from above and meets the required threshold value. Unlike the pure case, where the existence of an optimal solution is implied by the existence of a finite set of possible strategies, we have to provide an argument for the existence of an optimal strategy profile with given memory bound in this setting. According to the constraint system from above, the dictator assigns probabilities to the actions and selects the memory updates. If the resulting system complies with the first set of constraints, then it is a Nash resp. leader equilibrium. (Technically, ‘only if’ does not hold as these constraints only need to be satisfied by the reachable vertices. However, for the unreachable ones we can require the same without excluding relevant solutions.)

**Theorem 6.** For multi-player discounted sum games with perfect observation and predefined memory an optimal leader strategy profile exists.

**Proof.** First, we know that a strategy profile that satisfies the constraints exists (c.f., Section 6). Further, to see that an optimal strategy profile exists, we look at the reward obtained at the different probabilistic transitions. That is, we consider the reward obtained on the different probabilities assigned on different transitions. We define the payoff vector as a direct function on the probability assigned on the transitions and the strategy profile as the set $D$ (for decisions) of probability vectors over actions, or a finite dimensional closed subset of $[0, 1]^n$ for some $n \in \mathbb{N}$. This set of probability distributions over the possible actions gives the expected payoff for all players at all positions.
of the extended memory game graph (game graph with memory of pre-defined size $m$ at every vertex) and is defined by the memory copies at all vertices. The resultant payoff for all players at all vertices of the extended game graph is, thus, again a subset of a finite dimensional product of closed intervals, referred to as $\mathcal{P}$ (for payoff). The intervals are closed because, if $p$ defines the maximal absolute value of any of the individual payoffs in the discounted sum game, then every payoff must be in the closed interval $[-\frac{p}{1-\lambda}, \frac{p}{1-\lambda}]$. Given a strategy profile, represented by a $\vec{d} \in \mathcal{D}$, we can compute the payoffs, represented by a vector $\vec{p} \in \mathcal{P}$. We represent this by a valuation function $\text{val} : \mathcal{D} \to \mathcal{P}$, that maps each probability vector to a payoff vector. The valuation function is continuous: if the decision vector $\mathcal{D}$ changes only marginally, then the payoff vector $\mathcal{P}$ changes only marginally, too. Thus, if we fix an $\varepsilon > 0$ then we can first choose a natural number $l$, such that $\sum_{i=1}^{\infty} \lambda^i < \varepsilon$, and then choose a $\delta \in (0, 1]$ such that the change between two consecutive probabilities that is given by $l(1 - (1 - \delta)^i) < \frac{\varepsilon}{p}$ is only marginal. Then, if the absolute sum of changes of all probabilities is below $\delta$, we can estimate the difference by $\sum_{i=0}^{\infty} 2\lambda^i p (1 - (1 - \delta)^i)$. For the estimation of this difference, assume that we start with the probability vector $\vec{d}_m$, which is the pointwise minimum of $\vec{d}$ and $\vec{d}'$. Then the difference can be estimated by choosing the joint actions with the probability described in $\vec{d}_m$, and simply marking the positions with the missing probability (the difference between the sum of the probabilities reflected in $\vec{d}_m$ and 1 at every position in the extended game) as deviation. This difference is bounded by $\delta$.

The likelihood of being in a state where no difference has occurred so far is, after $i$ rounds, $\geq (1 - \delta)^i$. The likelihood that a difference has occurred so far can therefore be estimated by $(1 - (1 - \delta)^i)$. Using this estimation, we can estimate the difference, $\sum_{i=0}^{l-1} 2\lambda^i p (1 - (1 - \delta)^i) + \sum_{i=l}^{\infty} 2\lambda^i p (1 - (1 - \delta)^i) < \sum_{i=0}^{l-1} 2p(1 - (1 - \delta)^i) + 2\varepsilon < 4\varepsilon$, where the first inequality uses the definition of $l$, $\lambda^i \leq 1$, and $1 - (1 - \delta)^i < 1 - (1 - \delta)^l$, while the second estimation uses the definition of $\delta$. Thus, $\forall \varepsilon > 0 \exists \delta > 0$ such that $\| \vec{d} - \vec{d}' \| < \delta$ implies $\| \text{val}(\vec{d}) - \text{val}(\vec{d}') \| < 4\varepsilon$. The subset $\mathcal{C} \subseteq \mathcal{P}$ of the set of payoffs that comply with the constraint system is obviously still closed, as it is still a product of finitely many closed intervals. (Only the lower bound of these intervals may have changed.) As $\text{val}$ is continuous, the preimage $\mathcal{D}'$ of the closed set $\mathcal{C}$ is closed. When $\text{val}$ is restricted to $\mathcal{D}'$, then the maximum w.r.t. the value of the dictator in the initial state exists. That is, the supremum is taken for some value.

7 Conclusion

Starting from the observation that positional pure leader equilibria exist, we have inferred the existence of optimal positional and memory bounded pure strategies. We have argued that (and why) the detectability of deviation makes the restriction to pure strategies a natural choice. We have shown that the related decision problem (Is there a Nash resp. leader equilibrium that provides a payoff that meets or exceeds a given threshold?) is NP-complete. We also discuss the extension to mixed strategies and the extension of the observation model that is needed to make such strategies reasonable.
We have seen that, the nondeterministic approach for the construction of strategy profiles (which are Nash or leader equilibria and meet a given threshold for the overall payoff of the dictator) and the NP hardness proof, extends to mixed strategies in the extended observational model. Further, we have established the usefulness of memory. Unsurprisingly, more memory will always help. Unfortunately, our example from Figure 2 shows that there is no upper bound that can be inferred from the structure of the game such that one could know that memory of this size would suffice if unbounded memory would. Possible future work could be to implement the nondeterministic approach in SMT solvers like Yices [5,15] and see how they would fare on small examples.

References

[1] K. Berg and M. Kitti. Computing equilibria in discounted supergames. *Comput. Econ.*, 41(1):71–88, 2013.
[2] T. Bewley and E. Kohlberg. On stochastic games with stationary optimal strategies. *Mathematics of Operations Research*, 3(2):104–125, 1978.
[3] T. Brihaye, J. De Pril, and S. Schewe. Multiplayer cost games with simple Nash equilibria. In *Proc. of LFCS*, pages 59–73, 2013.
[4] K. Chatterjee, T. A. Henzinger, and M. Jurdzinski. Mean-payoff parity games. In *Proc. of LICS*, pages 178–187, 2005.
[5] L. M. de Moura, B. Dutertre, and N. Shankar. A tutorial on satisfiability modulo theories. In *Proc. of CAV*, pages 20–36, 2007.
[6] A. M. Fink. Equilibrium in a stochastic n-person game. 28:89–93, 1964.
[7] J. W. Friedman. A non-cooperative equilibrium for super-games. *The Review of Economic Studies*, 38(1):1–12, 1971.
[8] H. Gimbert and W. Zielonka. When can you play positionally? In *Proc. of MFCS*, pages 686–697, 2004.
[9] A. Gupta and S. Schewe. Quantitative Verification in Rational Environments. In *Proceedings of the 21st International Symposium on Temporal Representation and Reasoning*, pages 123–131, 2014.
[10] E. Lehrer. Nash equilibria of n-player repeated games with semi-standard information. *International Journal of Game Theory*, 19(2):191–217, 1990.
[11] J. F. Nash. Equilibrium points in n-person games. *Proc. of the National Academy of Sciences*, 36(1):48–49, 1950.
[12] M. J. Osborne and A. Rubinstein. *A course in game theory*. The MIT Press, 1994.
[13] L. I. Sennott. Zero-sum stochastic games with unbounded costs: Discounted and average cost cases. *Mathematical Methods of OR*, 39(2):209–225, 1994.
[14] L. S. Shapley. Stochastic Games. In *Proc. of the National Academy of Sciences*, volume 39, pages 1095–1100, 1953.
[15] Yices website: http://yices.csl.sri.com/.
[16] U. Zwick and M. S. Paterson. The complexity of mean payoff games on graphs. *TCS*, 158(1–2):343–359, 1996.