

Comparison of Algorithms for Simple Stochastic Games ^{*}

Jan Křetínský, Emanuel Ramneantu, Alexander Slivinskiy, Maximilian Weininger

Technical University of Munich

{jan.kretinsky,emanuel.ramneantu,alexander.slivinskiy,maxi.weininger}@tum.de

Simple stochastic games are turn-based 2½-player zero-sum graph games with a reachability objective. The problem is to compute the winning probability as well as the optimal strategies of both players. In this paper, we compare the three known classes of algorithms – value iteration, strategy iteration and quadratic programming – both theoretically and practically. Further, we suggest several improvements for all algorithms, including the first approach based on quadratic programming that avoids transforming the stochastic game to a stopping one. Our extensive experiments show that these improvements can lead to significant speed-ups. We implemented all algorithms in PRISM-games 3.0, thereby providing the first implementation of quadratic programming for solving simple stochastic games.

1 Introduction

Simple stochastic games (SGs), e.g. [16], are zero-sum games played on a graph by players Maximizer and Minimizer, who choose actions in their respective vertices (also called states). Each action is associated with a probability distribution determining the next state to move to. The objective of Maximizer is to maximize the probability of reaching a given target state; the objective of Minimizer is the opposite.

The basic decision problem is to determine whether Maximizer can ensure a reachability probability above a certain threshold if both players play optimally. This problem is among the rare and intriguing combinatorial problems that are in $\mathbf{NP} \cap \mathbf{co-NP}$ [15], but whether it belongs to \mathbf{P} is a major and long-standing open problem. Further, several other important problems can be reduced to SG, for instance parity games, mean-payoff games, discounted-payoff games and their stochastic extensions [7].

Besides the theoretical interest, SGs are a standard model in control and verification of stochastic reactive systems [20, 10]; see e.g. [39] for an overview over various recent case studies. Further, since Markov decision processes (MDP) [36] are a special case with only one player, SGs can serve as abstractions of large MDPs [28] or provide robust versions of MDPs when precise transition probabilities are not known [12, 40].

There are three classes of algorithms for computing the reachability probability in simple stochastic games, as surveyed in [15]: value iteration (VI), strategy iteration (SI, also known as policy iteration) and quadratic programming (QP). In [15], they all required the SG to be transformed into a certain normal form, among other properties ensuring that the game is stopping. This not only blows up the size of the SG, but also changes the reachability probability; however, it is possible to infer the original probability. For VI and SI, this requirement has since been lifted, e.g. [9, 6], but not for QP.

While searching for a polynomial algorithm, there were several papers on VI and SI; however, the theoretical improvements so far are limited to subexponential variants [35, 18] and variants that are fast for SG with few random vertices [21, 27], i.e. games where most actions yield a successor deterministically. QP was not looked at, as it is considered common knowledge that it performs badly in practice.

^{*}This research was funded in part by the German Research Foundation (DFG) projects 383882557 *Statistical Unbounded Verification (SUV)* and 427755713 *Group-By Objectives in Probabilistic Verification (GOPro)*.

There exist several tools for solving games: GAVS+ [14] offers VI and SI for SGs, among other things. However, it is more for educational purposes than large case studies and currently not maintained. GIST [11] performs qualitative analysis of stochastic games with ω -regular objectives. For MDPs (games with a single player), we refer to [23] for an overview of existing tools. Most importantly, PRISM-games 3.0 [34] is a recent tool that offers algorithms for several classes of games. However, for SGs with a reachability objective, it offers only variants of value iteration, none of which give a guarantee, so the result might be arbitrarily far off. Thus, currently no tool offers a precise solution method for large simple stochastic games.

Our contribution is the following:

- We provide an extension of the quadratic programming approach that does not require the transformation of the SG into a stopping SG in normal form.
- We propose several optimizations for the algorithms, inspired by [31], including an extension of topological value iteration from MDPs [19, 4].
- We implement VI with guarantees on precision as well as SI, QP and our optimizations in PRISM-games 3.0—thereby providing the first implementations of QP for SGs—and experimentally compare them on both the realistic case studies of PRISM-games and on interesting handcrafted corner cases.

Related Work

We sketch the recent developments of each class of algorithms and then several further directions.

Value iteration is a standard solution method also for MDPs [36]. For a long time, the stopping criterion of VI was such that it could return arbitrarily imprecise results [22]. In principle, the computation has to run for an exponential number of steps [9]. However, recently several heuristics that give guarantees and are typically fast were proposed for MDPs [5, 22, 37, 25], as well as for SGs [29]. A learning-based variant of VI for MDPs [5] was also extended to SGs [29]. The variant of VI from [27] requires the game to be in normal form, blowing up its size, and is good only if there are few random vertices, as it depends on the factorial of this number; it is impractical for almost all case studies considered in this paper.

Strategy iteration was introduced in [26]. A randomized version was given in [15], and subexponential versions in [35, 18]. Another variant of SI was proposed in [38], however there is no evidence that it runs in polynomial time. The idea of considering games with few random vertices (as in the already mentioned works of [18, 27]) was first introduced in [21], where they exhaustively search a subspace of strategies, which is called *f*-strategies.

To the best of our knowledge, quadratic programming as solution method for simple stochastic games was not investigated further after the first mention in [15]. However, convex QPs are solvable in polynomial time [30]. If it was possible to encode the problem in a convex QP of polynomial size, this would result in a polynomial algorithm. The encoding we provide in this paper can however be exponential in the size of the game.

Further related works consider other variants of the model, for example concurrent stochastic games, see e.g. [24] for the complexity of SI and VI and [8] for strategy complexity, or games with limited information [2]. Furthermore, one can consider other objectives, e.g. ω -regular objectives [10], mean payoff [41] or combinations of objectives [13, 1].

2 Preliminaries

We now introduce the model of stochastic games, define the semantics by the standard means of infinite paths and strategies and then define the important concept of end components, which are subgraphs of stochastic games that are problematic for all three classes of algorithms.

A *probability distribution* on a finite set X is a mapping $\delta : X \rightarrow [0, 1]$, such that $\sum_{x \in X} \delta(x) = 1$. The set of all probability distributions on X is denoted by $\mathcal{D}(X)$.

2.1 Stochastic Games

Now we define stochastic games, in literature often referred to as simple stochastic games or turn-based stochastic two-player games with a reachability objective. As opposed to the notation of e.g. [16], we do not have special stochastic nodes, but rather a probabilistic transition function.

Definition 1 (SG). A stochastic game (SG) is a tuple $(S, S_{\square}, S_{\circ}, s_0, A, Av, \delta)$ where

- S is a finite set of states partitioned¹ into the sets S_{\square} and S_{\circ} of states of the player Maximizer and Minimizer² respectively
- $s_0 \in S$ is the initial state
- A is a finite set of actions
- $Av : S \rightarrow 2^A$ assigns to every state a set of available actions
- $\delta : S \times A \rightarrow \mathcal{D}(S)$ is a transition function that given a state s and an action $a \in Av(s)$ yields a probability distribution over successor states. We slightly abuse notation and write $\delta(s, a, s')$ instead of $\delta(s, a)(s')$.

See Figure 1 for an example of an SG. Since we consider the reachability objective, the SG is complemented with a set of target states $F \subseteq S$. What happens after reaching any target state is irrelevant for the reachability probability, so we can assume that every target state only has one action that is a self-loop with probability 1. A Markov decision process (MDP) is a special case of SG where $S_{\circ} = \emptyset$, and a Markov chain (MC) is a special case of an MDP, where for all $s \in S : |Av(s)| = 1$.

Without loss of generality we assume that SGs are non-blocking, so for all states s we have $Av(s) \neq \emptyset$. For a state s and an available action $a \in Av(s)$, we denote the set of successors by $Post(s, a) := \{s' \mid \delta(s, a, s') > 0\}$. Finally, for any set of states $T \subseteq S$, we use T_{\square} and T_{\circ} to denote the states in T that belong to Maximizer and Minimizer, whose states are drawn in the figures as \square and \circ , respectively.

2.2 Semantics: Paths, Strategies and Values

The semantics of SGs is given in the usual way by means of strategies, the induced Markov chain and the respective probability space, as follows: An *infinite path* ρ is an infinite sequence $\rho = s_0 a_0 s_1 a_1 \dots \in (S \times A)^{\omega}$, such that for every $i \in \mathbb{N}$ we have $a_i \in Av(s_i)$ and $s_{i+1} \in Post(s_i, a_i)$. *Finite paths* are defined analogously as elements of $(S \times A)^* \times S$.

As this paper deals with the reachability objective, we can restrict our attention to memoryless deterministic strategies, which are optimal for this objective [16]. A *strategy* of Maximizer, respectively

¹I.e., $S_{\square} \subseteq S$, $S_{\circ} \subseteq S$, $S_{\square} \cup S_{\circ} = S$, and $S_{\square} \cap S_{\circ} = \emptyset$.

²The names are chosen, because Maximizer maximizes the probability of reaching the given target states, and Minimizer minimizes it.

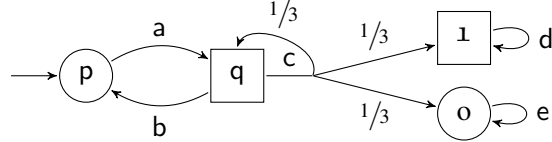


Figure 1: An example of an SG with $S = \{p, q, 1, o\}$, $S_{\square} = \{q, 1\}$, $S_{\circ} = \{p, o\}$, the initial state p and the set of actions $A = \{a, b, c, d, e\}$; $Av(p) = \{a\}$ with $\delta(p, a)(q) = 1$; $Av(q) = \{b, c\}$ with $\delta(q, b)(p) = 1$ and $\delta(q, c)(q) = \delta(q, c)(1) = \delta(q, c)(o) = \frac{1}{3}$. For actions with only one successor, we do not depict the transition probability 1.

Minimizer, is a function $\sigma : S_{\square} \rightarrow A$, respectively $S_{\circ} \rightarrow A$, such that $\sigma(s) \in Av(s)$ for all s . A pair (σ, τ) of memoryless deterministic strategies of Maximizer and Minimizer induces a Markov chain $G^{\sigma, \tau}$, as for every state there is only one available action. The Markov chain induces a unique probability distribution $\mathbb{P}_s^{\sigma, \tau}$ over measurable sets of infinite paths [3, Ch. 10].

We write $\diamond F := \{\rho \mid \rho = s_0 a_0 s_1 a_1 \dots \in (S \times A)^{\omega} \wedge \exists i \in \mathbb{N}. s_i \in F\}$ to denote the (measurable) set of all paths which eventually reach F . For each $s \in S$, we define the *value* in s as

$$V(s) := \sup_{\sigma} \inf_{\tau} \mathbb{P}_s^{\sigma, \tau}(\diamond F) = \inf_{\tau} \sup_{\sigma} \mathbb{P}_s^{\sigma, \tau}(\diamond F),$$

where the equality follows from [16]. The value is the least fixpoint of the so called *Bellman equations* [15]:

$$V(s) = \begin{cases} 1 & \text{if } s \in F \\ \max_{a \in Av(s)} V(s, a) & \text{if } s \in S_{\square} \setminus F, \\ \min_{a \in Av(s)} V(s, a) & \text{if } s \in S_{\circ} \setminus F \end{cases}, \quad (1)$$

$$\text{with } V(s, a) := \sum_{s' \in S} \delta(s, a, s') \cdot V(s') \quad (2)$$

The states with no path to the target, so called sinks, are of special interest. We denote the set of sinks as Z . Sinks have a value of 0 and can be found a priori by graph analysis.

We are interested not only in the values $V(s)$ for all $s \in S$, but also their ε -approximation, i.e. an approximation L with $|V(s) - L(s)| < \varepsilon$; as well as the corresponding (ε -)optimal strategies for both players, i.e. a pair of strategies (σ, τ) under which $\mathbb{P}_s^{\sigma, \tau}(\diamond F)$ is equal to $V(s)$ (respectively an ε -approximation of $V(s)$). Note that it suffices to have either the values or the optimal strategies, because from one we can infer the other. Given a pair of optimal strategies (σ, τ) , the values can be computed by solving the induced Markov chain $G^{\sigma, \tau}$. Given a vector of values for all states, an optimal pair of strategies can be computed by randomizing over all locally optimal actions in each state (e.g. $\sigma(s)$ randomizes over the set $\arg \max_{a \in Av(s)} V(s, a)$ for Maximizer states, and dually with $\arg \min$ for Minimizer). To get a deterministic strategy, we cannot only pick some locally optimal action, but additionally we have to ensure that Maximizer is not stuck in some cycle when playing the actions. This works by looking at end components, which are the topic of the next subsection.

2.3 End Components

When computing the values of states in an SG, we need to take special care of *end components* (EC). Intuitively, an EC is a subset of states of the SG, where the game can remain forever; i.e. given certain strategies of both players, there is no positive probability to exit the EC to some other state. ECs

correspond to bottom strongly connected components of the Markov chains induced by some pair of strategies.

Definition 2 (EC). A non-empty set $T \subseteq S$ of states is an end component (EC) if there exists a non-empty set $B \subseteq \bigcup_{s \in T} \text{Av}(s)$ of actions³ such that

1. for each $s \in T, a \in B \cap \text{Av}(s)$ we have $\text{Post}(s, a) \subseteq T$,
2. for each $s, s' \in T$ there is a finite path $w = sa_0 \dots a_n s' \in (T \times B)^* \times T$, i.e. the path stays inside T and only uses actions in B .

An end component T is a maximal end component (MEC) if there is no other end component T' such that $T \subseteq T'$.

Given an SG G , the set of its MECs is denoted by $\text{MEC}(G)$ and can be computed in polynomial time [17].

Example 1. Consider the SG of Figure 1. The set of states $T = \{p, q\}$ is an EC, as when playing only actions from $B = \{a, b\}$ the play remains in T forever. It is even a MEC, as there is no superset of T with this property. \triangle

ECs are of special interest, because in ECs there can be multiple fixpoints of the Bellman equations (see Equation 1 and 2). Thus, methods relying on iterative over-approximation of the value do not converge, as they are stuck at some greater fixpoint than the value (cf. [29, Lemma 1] and Example 2 in the next section). This is why the original description of the algorithms [15] considered only stopping games, i.e. games without ECs, except for a dedicated target and sink state. The algorithms are theoretically applicable to arbitrary SGs, as for every non-stopping SG one can construct a stopping SG and infer the original value from solving the stopping SG. See [33, Appendix A.4] for a description of this approach and Section 4.1.4 for a discussion of the practical drawbacks.

3 State of the Art Algorithms

In this section, we describe the existing algorithms for solving simple stochastic games. All of them require as input an SG $G = (S, S_{\square}, S_{\circ}, s_0, A, \text{Av}, \delta)$ and a target set F . Value iteration additionally needs a precision ε . After termination, all of them return a vector of values for each state (ε -precise for BVI) and the corresponding (ε -)optimal strategies. Quadratic programming in its current form only works on stopping SGs in a certain normal form.

3.1 Bounded Value Iteration

Value Iteration (VI), see e.g. [36], is the most common algorithm for solving MDPs and SGs, and the only method implemented in PRISM-games [34]. Originally, VI only computed a convergent sequence of under-approximations; however, as it was unclear how to stop, results returned by model checkers could be off by arbitrary amounts [22]. Thus, it was extended to also compute a convergent over-approximation [29]. The resulting algorithm is called *bounded value iteration (BVI)*.

The basic idea of BVI is to start from a vector L_0 respectively U_0 that definitely is an under-/over-approximation of the value, i.e. for every state $L_0(s) \leq V(s) \leq U_0(s)$. Then the algorithm repeatedly

³Note that this assumes that action names are unique. This can always be achieved by renaming actions, e.g. prepending every action with the state it is played from

applies so called *Bellman updates*, i.e. it uses a version of Equation 1 as follows (the equation for the over-approximation is obtained by replacing L with U):

$$L_n(s) = \begin{cases} \max_{a \in \text{Av}(s)} L_{n-1}(s, a) & \text{if } s \in S_{\square} \\ \min_{a \in \text{Av}(s)} L_{n-1}(s, a) & \text{if } s \in S_{\circ} \end{cases}, \quad (3)$$

where $L_{n-1}(s, a)$ is computed from $L_{n-1}(s)$ as in Equation 2. Since V is the least fixpoint of the Bellman equations, $\lim_{n \rightarrow \infty} L_n$ converges to V . However, the over-approximation U need not converge in the presence of ECs.

Example 2. Consider the SG of Figure 1 with the EC $\{p, q\}$. Let $U_0 = 1$ for p, q and \perp and $U_0 = 0$ for o . Then we have $U_0(q, b) = 1$ and $U_0(q, c) = 2/3$. Thus, q will pick action b , as it promises a higher value, and the over-approximation does not change. This happens, because looking at the current upper bound, Maximizer is under the impression that staying in the EC yields a higher value. However, it actually reduces the probability to reach the target to 0. So the algorithm has to perform an additional step to inform states in an EC that they should not depend on each other, but on the best exit. In this case, $U_1(q) = U_0(q, c) = 2/3$. \triangle

In fact, it does not suffice to only decrease the upper bound of ECs, but a more in-depth graph analysis is required to detect the problematic subsets of states, so called *simple end components* (SEC) [29]. Decreasing the value of those SECs to their best exit repeatedly results in a convergent over-approximation.

$$\text{best_exit}_U(T) = \max_{\substack{s \in T_{\square} \\ \neg \text{Post}(s, a) \subseteq T}} U(s, a)$$

is the best exit according to the current estimates of the upper bound. Algorithm 1 shows the full BVI algorithm from [29].

Algorithm 1 Bounded value iteration algorithm from [29]

```

1: procedure BVI(precision  $\varepsilon > 0$ )
2:   for  $s \in S$  do      # Initialization
3:      $L(s) = 0$       # Lower bound
4:      $U(s) = 1$       # Upper bound
5:   for  $s \in F$  do  $L(s) = 1$     # Value of targets is determined a priori
6:   repeat
7:      $L, U$  get updated according to Eq. (3)    # Bellman updates
8:     for  $T \in \text{FIND\_SEC}$  do      # For every SEC
9:       for  $s \in T$  do
10:         $U(s) \leftarrow \text{best\_exit}_U(T)$     # Decrease upper bound to best exit
11:   until  $U(s) - L(s) < \varepsilon$  for all  $s \in S$     # Guaranteed error bound

```

There also is a simulation based asynchronous version of BVI (see [29, Section 4.4]) which can perform very well on models with a certain structure [32]. It updates states encountered by simulations, and guides those simulations to explore only the relevant part of the state space. If only few states are relevant for convergence, this algorithm is fast; if large parts of the state space are relevant or there are many cycles in the game graph that slow the simulations down, this adaption of BVI is slow.

3.2 Strategy Iteration

In contrast to value iteration, the approach of *strategy iteration (SI)* [26] does not compute a sequence of value-vectors, but instead a sequence of strategies. Starting from an arbitrary strategy of Maximizer, we repeatedly compute the best response of Minimizer and then greedily improve Maximizer's strategy. The resulting sequence of Maximizer strategies is monotonic and converges to the optimal strategy [6, Theorem 3]. The pseudocode for strategy iteration is given in Algorithm 2.

Note that in non-stopping SGs (games with ECs) the initial Maximizer strategy cannot be completely arbitrary, but it has to be *proper*, i.e. ensure that either a target or a sink state is reached almost surely; it must not stay in some EC, as otherwise the algorithm might not converge to the optimum due to problems similar to those described in Example 2. In Algorithm 2, we use the construction of the *attractor strategy* [6, Section 5.3] to ensure that our initial guess is a proper strategy (Line 2). It first analyses the game graph to find the sink states Z . Then, it performs a backwards breadth first search, starting from the set of both target and sink states. A state discovered in the i -th iteration of the search has to choose some action that reaches a state discovered in the $(i-1)$ -th iteration with positive probability. Such a state exists by construction, and the choice ensures that the initial strategy reaches the set of target or sink states almost surely.

When a Maximizer strategy is fixed, the main loop of Algorithm 2 solves the induced MDP G^σ (Line 6). It need not remember the Minimizer strategy, but only uses the computed value estimates L to greedily update Maximizer's strategy (Line 8); note that here $L(s, a)$ is again computed from $L(s)$ as in Equation 2. The algorithm stops when the Maximizer strategy does not change any more in one iteration. We can then compute the values and the corresponding Minimizer strategy by solving the induced MDP G^σ .

Algorithm 2 Strategy iteration

```

1: procedure SI
2:    $\sigma' \leftarrow$  arbitrary Maximizer attractor strategy    # Proper initial strategy
3:   repeat
4:      $\sigma \leftarrow \sigma'$ 
5:     for  $s \in S$  do
6:        $L(s) \leftarrow \inf_{\tau} \mathbb{P}_s^{\sigma, \tau}(\diamond F)$           # Solve MDP for opponent
7:     for  $s \in S_{\square}$  do
8:        $\sigma'(s) \leftarrow \arg \max_{a \in Av(s)} L(s, a)$     # Greedily optimise choices
9:   until  $\sigma = \sigma'$ 

```

3.3 Quadratic Programming

Quadratic programming (QP) [15] works by encoding the graph of the SG in a system of constraints. The only global (and local) optimum of the objective function under these constraints is 0, and it is attained if and only if the variable for every state is set to the value of that state [15, Section 3.1]. The proof as well as the construction of the quadratic program relies on the game being in a certain normal form, which consists of four conditions:

- **2Act:** For all $s \in S$: $|Av(s)| \leq 2$.
- **No1Act:** If $|Av(s)| = 1$, then s is a target or a sink.

- **1/2Probs:** For all $s, s' \in S, a \in \text{Av}(s) : \delta(s, a, s') \in \{0, 0.5, 1\}$.
- **Stopping game:** There are no ECs in G (except for the sinks and targets).

We shortly discuss the advantage of each condition of the normal form: the reason for **2Act** and **No1Act** is that the objective function of the QP requires every (non-sink and non-target) state to have exactly 2 successors. Arguing about a game with average nodes instead of actions mapping to arbitrary probability distribution simplified the proofs, which is the advantage of **1/2Probs**. **Stopping game** was necessary, because of the problem of non-convergence in end components, as in Example 2. We recall the procedure from [16] to transfer an arbitrary SG into a polynomially larger SG in normal form in the full version [33, Appendix A].

We now state the quadratic program for an SG in normal form as given in [15], but adjusted to our notation. Intuitively, the objective function is 0 if all summands are 0. And the summand for some state s is 0 if its value $V(s)$ is equal to the value of one of its actions $V(s, a)$ or $V(s, b)$. The program is quadratic, since all states (except targets and sinks) have exactly two successors and hence the summand for every state is a quadratic term. The constraints encode the game, ensuring that Maximizer/Minimizer states use the action with the highest/lowest value and fixing the values of targets and sinks. Note the additional definition of $V(s, a)$, which assumes that all occurring non-trivial probabilities are $1/2$.

$$\begin{array}{ll}
\text{minimize} & \sum_{s \in S} (V(s) - V(s, a))(V(s) - V(s, b)) \\
\text{subject to} & \begin{array}{ll}
\text{Av}(s) = \{a, b\} & \\
V(s) \geq V(s, a) & \forall s \in S_{\square} : |\text{Av}(s)| = 2, \forall a \in \text{Av}(s) \\
V(s) \leq V(s, a) & \forall s \in S_{\circ} : |\text{Av}(s)| = 2, \forall a \in \text{Av}(s) \\
V(s) = 1 & \forall s \in F \\
V(s) = 0 & \forall s \in Z
\end{array} \\
\text{where } V(s, a) = & \left\{ \begin{array}{ll}
V(s') & \text{for } |\text{Post}(s, a)| = \{s'\} \\
1/2V(s') + 1/2V(s'') & \text{for } |\text{Post}(s, a)| = \{s', s''\}
\end{array} \right\}
\end{array}$$

4 Improvements

In this section, we first generalize QP to be applicable to arbitrary SGs, thereby omitting the costly transformations into the normal form. Then, we identify a hyperparameter of SI. Finally, we provide two optimizations that are applicable to all three algorithms.

4.1 Quadratic Programming for General Stochastic Games

Every transformation into the normal form (see [33, Appendix A]) adds additional states or actions to the SG. Firstly, we want to change the constraints of the QP so that it can deal with arbitrary SGs; secondly, we want to avoid blowing up the SG, as the time for solving the QP depends on the size of the SG.

4.1.1 2Act

In order to drop the constraint that every state has at most two actions, we can generalize the summand of a state s in the objective function to $\prod_{a \in \text{Av}(s)} (V(s) - V(s, a))$. The resulting program is no longer quadratic, as for a state with n actions now the objective function has order n . Thus, in the experiments,

we report the verification times of both (i) a higher-order optimization problem for the original SG as well as (ii) a QP for the SG that was transformed to comply with **2Act**.

One more step is needed to ensure that the objective function still is correct: Recall that we want the only global optimum of the objective function to be 0, and it should be attained if and only if $V(s) = V(s, a)$ for some action $a \in Av(s)$. However, if a Minimizer state has an odd number of actions, its summand in the objective function could be negative. For example, a Minimizer state with three actions could choose its value to be smaller than all three actions. Multiplying three negative numbers results in a negative number, which is preferred to a summand of 0, since we minimize the objective function. Thus, for a state with an odd number of successors, we duplicate the term for one of the actions. Then the summand for every state is non-negative. For a Maximizer state, all factors are greater or equal to 0, since by the first constraint of the QP we have $V(s) \geq V(s, a)$ for every action. For a Minimizer state, either one of the factors is 0 or all factors are negative (by the second constraint of the QP), and multiplying an even number of negative numbers results in a positive number. Thus, 0 is the only global optimum of the objective function in the constrained region.

4.1.2 No1Act

The **No1Act**-constraint compels every state $s \in S$ that is not a target or a sink to have more than one action. The transformation for complying with this constraint (see [33, Appendix A.2]) adds a second action to every state with only one action; however, the newly added action is chosen in such a way that it does not influence the value of the state, and can thus also be omitted (see [33, Appendix B.1] for the formal argument). For a non-absorbing state s with only one action a , we can simplify the program by not including it in the objective function, but only adding a single constraint $V(s) = V(s, a)$.

4.1.3 1/2Probs

To comply with the **1/2Probs**-requirement, every transition $\delta(s, a, s')$ with $s, s' \in S$ and $a \in Av(s)$ must have a probability of 0, 0.5 or 1. We can omit this constraint if we use the general definition of $V(s, a)$ as in Equation 2, summing the successors with their respective given transition probability (see [33, Appendix B.2] for the formal argument).

4.1.4 Stopping Game

The final and most complicated constraint of the normal form is that we require the SG to be stopping. The transformation of an arbitrary game into a stopping one (see [33, Appendix A.4]) adds a transition to a sink with a small probability ε to every action, thus ensuring that a sink (or a target) is reached almost surely. This is problematic not only because the added transitions blow up the quadratic program, but even more so because of the following:

The ε -transitions modify the value of the game. Theoretically, this is no problem, because the value is rational and we know the greatest possible denominator q it can have. Thus, by choosing ε sufficiently small, we ensure that the modified value does not differ by more than $1/q$ and we can obtain the original value by rounding. However, practically, this denominator becomes smaller than machine precision even for small systems, resulting in immense numerical errors. The ε has to be strictly smaller⁴ than

⁴It actually has to be a lot smaller, since the *value* of every state may differ by at most that amount, but this conservative upper bound suffices to prove our point.

$(1/4)^{|S|}$ [16]. According to IEE 754.2019 standard⁵, the commonly used double machine precision is 10^{-16} , so already for 27 states the necessary ε becomes smaller than machine precision. Thus, the transformation to a stopping game is inherently impractical.

Our approach introduces additional constraints for every maximal end component (MEC) to ensure that the QP finds the correct solution.

For MECs where all states belong to the same player, the solution is straightforward. All states in MECs with only Minimizer states have a value of 0, as they can choose to remain forever in the EC and not reach the target; they can be identified a priori and are part of the set of sinks Z . All states in MECs with only Maximizer states have the value of the best exit from that MEC [29]. Thus, for all $T \in \text{MEC}(\mathcal{G})$ with $T \cap S_{\bigcirc} = \emptyset$ we can introduce an additional constraint: $\forall s \in T : V(s) = \text{best_exit}_V(T)$, where best_exit is defined as for BVI (see Section 3.1). Note that max-constraints are expressed through continuous and boolean variables and therefore, the resulting quadratic program is a mixed-integer program.

For MECs containing states of both players, the values of the states depend on the best exit that Maximizer can ensure reaching against the optimal strategy of Minimizer. We cannot just set the value to the best exit of the whole MEC, as Minimizer might prevent some states in the MEC from reaching that best exit. The solution of [29] to analyse the graph and figure out Minimizer's decisions on the fly is not possible, because we have to give the constraints a priori.

We solve the problem as follows: iterate over all strategy-pairs (σ, τ) in the MEC and for each pair describe the corresponding value of every state $V_{\sigma, \tau}(s)$ depending on the values of the exiting actions $E^T = \{(s, a) \mid s \in T \wedge \text{Post}(s, a) \not\subseteq T\}$. Then constrain the value for all states in the MEC according to the optimal strategies, i.e. $V(s) = \max_{\sigma} \min_{\tau} V_{\sigma, \tau}(s)$. This ensures that the value of every state is set to the best exit it can reach, because the optimal strategies are chosen.

It remains to define how to describe $V_{\sigma, \tau}$ depending on the exits E^T . For a pair of strategies (σ, τ) in the MEC, we consider the induced Markov chain $G^{\sigma, \tau}$. We modify the MC and let every state-action pair $(s, a) \in E^T$ lead to a sink state $e_{(s, a)}$. Then, for every such sink state, we compute the probability $p_{e_{(s, a)}}^{\sigma, \tau}(t)$ to reach it from every state $t \in T$ by solving the MC. Then we set $V_{\sigma, \tau}(t) = \sum_{(s, a) \in E^T} p_{e_{(s, a)}}^{\sigma, \tau}(t) \cdot V(s, a)$. We summarize the procedure we just described in Algorithm 3. We use *all strategies on T* to denote every possible mapping that maps every state $t \in T$ to some available action $a \in \text{Av}(t)$.

Algorithm 3 Algorithm to add constraints for MECs containing states of both players

```

1: procedure ADD_MEC_CONSTRAINTS(MEC  $T \subseteq S$ )
2:   for all strategies  $\sigma$  on  $T_{\square}$  do
3:     for all strategies  $\tau$  on  $T_{\bigcirc}$  do
4:       for every  $(s, a) \in E^T$  do
5:         Compute  $p_{e_{(s, a)}}^{\sigma, \tau}(t)$  for all  $t \in T$  by solving the modified induced MC  $G^{\sigma, \tau}$ 
6:       for every  $t \in T$  do
7:         Add constraint:  $V(t) = \max_{\sigma} \min_{\tau} \sum_{(s, a) \in E^T} p_{e_{(s, a)}}^{\sigma, \tau}(t) \cdot V(s, a)$ 

```

Note that $V(s, a)$ is defined as usual (see Equation 2). As by definition of being an exit it depends on some successor $s' \notin T$, the states in the MEC cannot depend only on each other any more, but they

⁵<https://standards.ieee.org/content/ieee-standards/en/standard/754-2019.html>

have to depend on exiting actions. This ensures there is a unique solution. See [33, Appendix B.3] for the formal proof.

For a MEC of size n we have to examine at most $n^{(\max_{s \in T} |Av(s)|)}$ pairs of strategies, because it suffices to consider memoryless deterministic strategies. Since the min and max constraints have to be explicitly encoded, we have to add a number of constraints that is exponential in the size of the MEC.

In summary, we have shown how to replace every condition of the normal form by modifying the constraints and objective function of the program. The modifications always ensure that the program still computes the correct value, because it still only has a single global optimum in the constrained region, namely if every state variable is set to its value. Thus, we can provide a higher-order program to solve arbitrary SGs, and a quadratic program to solve SGs that only have to satisfy the **2Act** condition.

4.2 Opponent Strategy for Strategy Iteration

We can tune the MDP solution method that is used to compute the opponent strategy in Line 6 of Algorithm 2. We need to ensure that we fix the new choices of Maximizer correctly. For this, we can use the precise MDP solution methods strategy iteration or linear programming (LP, a QP with an objective function of order 1). However, we do not need the precise solution of the induced MDP, but it suffices to know that an action is better than all others. So we can also use bounded value iteration and check that the lower bound of one action is larger than the upper bound of all other actions, and thus we can stop the algorithm earlier. This approximation and the fact that VI tends to be the fastest methods in MDPs can speed up the solving. Using unguaranteed VI is dangerous, as it might stop too early and return a wrong strategy.

4.3 Warm Start

All three solution methods can benefit from prior knowledge. VI and quadratic/higher-order programs (QP/HOP) can immediately use initial solution vectors obtained by domain knowledge or any precomputation. For VI, it is necessary to know whether it is an upper or lower estimate to use the prior knowledge correctly. The QP/HOP optimization process can start at the given initial vector.

SI can use the information of an initial estimate to infer a good initial strategy, as already suggested in [31]. This reduces the number of iterations of the main loop and thus the runtime. However, we have to ensure that the resulting strategy is proper. For example, we can check whether the target and sink states are reached almost surely from every state, and if not, we change the strategy to an attractor strategy where necessary, preferring those allowed actions that have a higher value.

We can also improve the MDP-solving for SI (Line 6 of Algorithm 2) by giving it the knowledge we currently have. We anyway save the estimate L of the previous iteration and, since the strategies of Maximizer get monotonically better, L certainly is a lower bound for the values in the MDP.

Even in the absence of domain knowledge or sophisticated precomputations, we can run unguaranteed VI first in order to get some estimates of the values. Then we can use those estimates for SI and QP/HOP. Note that this is similar in spirit to the idea of optimistic value iteration [25]: utilize VI's ability to usually deliver tight lower bounds and then verify them.

4.4 Topological Improvement

For MDPs, topological improvements have been proposed for VI [19] and for SI [31, Algorithm 3]. These utilize the fact that the underlying graph of the MDP can be decomposed into a directed acyclic

graph of strongly connected components (SCC). Intuitively speaking, there are parts of the graph that, after leaving them, can never be reached again. Their value depends solely on the parts of the state space that come after them. So instead of computing the values on the whole game at once, one can iterate over the SCC in a backwards fashion, starting with the target and sink states and then propagating the information and solving the SCCs one by one according to their topological ordering.

This idea was extended to BVI for MDPs in [4]. The proof generalizes to SGs, as the basic argument of the topological ordering of SCCs is independent from introducing a second player. As the proof relies on the solutions for the later components being ϵ -precise, solving those components with the precise methods SI or QP is of course also possible.

5 Experimental Results

Implementation: We implemented all our algorithms as an extension of PRISM-games 3.0 [34]. They are available via github <https://github.com/ga67vib/Algorithms-For-Stochastic-Games>. For BVI, we reimplemented the algorithm as described in [29]; for SI and QP, this is the first implementation in PRISM-games. To solve the quadratic program, we used Gurobi⁶ or CPLEX⁷. For the higher-order programs, we constructed them with AMPL and solved them with MINOS⁸.

Setup: All experiments were conducted on a Linux Manjaro server with a 3.60 GHz Intel(R) Xeon(R) W-2123 CPU and 64 GB of RAM. We used a timeout of 15 minutes and set the java heap size for PRISM-games to 32 GB and the stack size to 16 GB⁹. The precision for BVI was set to 10^{-6} . In theory, the other algorithms are precise; in practice, QP and higher-order programming can have numerical problems. For a discussion about the practical precision of SI we refer to the end of Section 5.4 as well as [33, Appendix C.3].

Case studies: As case studies, we used those that are distributed with PRISM-games 3.0 as well as some handcrafted corner cases; see [33, Appendix C.1] for a detailed description.

We first analyze the impact of our optimizations by comparing different variants of the same algorithm. The full tables that this analysis is based on are in [33, Appendix C.2], but general trends are also visible in Table 1. Based on this, we select the best variants of each algorithm and compare between the algorithms.

5.1 Value Iteration

We could reproduce most of the findings of [29]: the overhead of BVI compared to unguaranteed VI is usually negligible and not performing the expensive deflate operation in every step speeds up the computation. Unguaranteed VI fails on three of the models. In contrast to [29], we found no model where the simulation based asynchronous version BRTDP was significantly faster than BVI. In fact, BRTDP is only faster on a single model (cloud6, where BVI takes 2 seconds), but significantly slower on many others, often even failing to produce results in time. Note that the implementation of BRTDP was in PRISM-games 2, and thus the disadvantage may also have technical reasons; improvements in the simulation engine or data structures might lead to speed-ups that make BRTDP competitive again.

The new topological variant of BVI (called TBVI) is usually in the same order of magnitude as the default approach. The exception to this are the models AV15_15 and especially MulMec, where TBVI

⁶<https://www.gurobi.com/>

⁷<https://www.ibm.com/analytics/cplex-optimizer>

⁸<https://ampl.com/products/solvers/solvers-we-sell/minos/>

⁹-javamaxmem 32g -javastack 16g

is a lot slower. A possible explanation is that TBVI solves every SCC of the model with a precision of ϵ . So when one SCC is solved and has a difference of almost exactly ϵ between upper and lower bound, SCCs depending on it take a longer time to converge, as the information about their exits is suboptimal. This is particularly problematic when the model has a structure like MulMec, namely a chain of MECs, and hence a chain of SCCs. This problem is not specific to topological VI for SGs, but can also occur for MDPs.

5.2 Strategy Iteration

Using BVI for the opponent's MDP and the warm start usually lead to small speed-ups. We did not consider using linear programming for the opponent's MDP, as it is not supported by PRISM. The topological variant is significantly better in `two_inv` and `MulMec_e3`, but on the rest of the models performs very similar to the non-topological version. Combining topological SI and BVI for the opponent's MDP leads to the same problems with MulMec as when using topological BVI. In contrast, topological SI with SI for the opponent's MDP works, because SI solves the SCCs precisely, allowing the SCCs depending on the previous solutions to converge as well.

5.3 Quadratic Programming

The original version of quadratic programming [15], that requires to transform the SG into normal form is impractical, producing a result within 15 minutes for only 3 of 34 case studies, and even there taking a lot more time than the improved version. After dropping the constraints **No1Act** and **1/2Probs**, it can correctly solve 14 of the case studies in time. Both of these variants are prone to the numerical errors described in Section 4.1.4 and produce incorrect results on some models. The quadratic program obtained after dropping all but the **2Act** constraint is solved successfully by Gurobi in 19 instances; CPLEX on the other hand only solves 7 instances correctly, one time even reporting an incorrect result. The warm start helps Gurobi on the model `charlton1`. Several other times, Gurobi discards the given initial suggestion and uses its own heuristics; thus, the warm start sometimes incurs a slight overhead.

Dropping all constraints, the higher-order program (HOP) with the solver Minos gets the correct result on 23 instances. The HOP is typically faster than the QP, except on the models `HW` and `AV`. The topological variant of both the quadratic programs as well as the higher-order program can lead to significant speed-ups, for example on `charlton1`, `mdsm1`, `two_inv` or `HW10_10_2`. The topological HOP is strictly better than all other algorithms in this subsection.

To estimate the impact of the EC solution method, we used several handcrafted or modified models: A single large MEC (`BigMec_e2`, with a MEC of size 201) cannot be solved with our approach, as there are too many choices; possibly, some heuristic could help identify reasonable strategies. In contrast, small MECs do not affect runtime a lot. The models `cdmsn` and `dice50` prepended with a single three-state MEC are solved in the same time as the original models. Even a chain of 1000 three-state MECs can be solved quickly (`MulMec_e3`).

5.4 Comparison

Comparing the algorithms, we see that BVI and SI perform very similar. BVI succeeds in the largest version of `BigMec`, but SI is the only one to solve the large and complicated models `AV15_15_2/3`, and TSI has the best runtime in `MulMec_e3`. TSI benefits from a model with small subcomponents that it can

Table 1: Table for the verification times (in seconds) of several variations of the algorithms. An X in the table denotes that the computation did not finish within 15 minutes. A red background colour indicates that the returned result was wrong. The four left-most columns give information about case study: its name, its size, the maximum/average number of actions and the number of relevant MECs. This table shows a selection of the most interesting case studies. They are roughly sorted by increasing size/difficulty, with scaled versions of the same model grouped together. The considered algorithms are the default and topological variant of BVI (with deflating only happening every 100 iterations); SI with BVI as MDP solver and topological SI with warm start and SI as MDP solver; and QP with Gurobi as solver and warm start, as well as the topological higher order program.

Case Study	States	Acts	MECs	BVI ₁₀₀	TBVI ₁₀₀	SI _{BV}	TSI _{SI} ^W	QP _G ^W	THOP
prison_dil	102	3/1.34	0	<1	<1	<1	<1	8	<1
charlton1	502	3/1.56	0	<1	<1	<1	<1	144	<1
cdmsn	1,240	2/1.66	0	<1	<1	<1	<1	<1	<1
cdmsnMec	1,244	2/1.66	1	<1	<1	<1	<1	<1	<1
cloud6	34,954	13/4.45	2176	2	3	<1	<1	X	X
mdsn1	62,245	2/1.34	0	5	3	5	3	X	3
dice50	96,295	2/1.48	0	6	6	6	6	X	6
dice50Mec	96,299	2/1.48	1	6	6	7	6	X	6
two_inv	172,240	3/1.34	0	13	13	19	13	X	20
HW10_10_1	400,000	5/2.52	0	10	11	11	11	98	11
HW10_10_2	400,000	5/2.52	0	<1	1	2	2	49	1
AV10_10_1	106,524	6/2.17	0	<1	<1	<1	<1	3	<1
AV10_10_2	106,524	6/2.17	6	72	70	79	77	X	X
AV10_10_3	106,524	6/2.17	1	45	55	50	60	X	X
AV15_15_1	480,464	6/2.14	0	1	1	2	2	11	2
AV15_15_2	480,464	6/2.14	6	X	X	825	X	X	X
AV15_15_3	480,464	6/2.14	1	X	X	500	X	X	X
hm_30	61	1/1.00	0	X	X	X	X	<1	<1
MulMec_e2	302	2/1.99	100	2	X	X	<1	<1	<1
MulMec_e3	3,002	2/2.00	1000	151	X	X	3	7	4
BigMec_e2	203	2/1.99	1	<1	<1	<1	<1	X	X
BigMec_e3	2,003	2/2.00	1	1	1	1	1	X	X
BigMec_e4	20,003	2/2.00	1	226	230	X	X	X	X

quickly solve (as in MulMec), while BVI is fast in subgraphs without probabilistic cycles (as the large chains in BigMec). Depending on the model structure, both of these algorithms are a viable choice.

Topological higher-order programming (THOP), the improved version of QP, is comparable on many case studies, but still a lot worse on several others, e.g. cloud6 and BigMec. This volatility is even more pronounced for the QP. The reason for this can be MECs which blow up the QP, but it can also happen in models with few or no MECs; in the latter case, we do not know which property of the model slows down the solving. Note that QP and HOP are able to solve models with many small MECs (MulMec_e3) quickly, while already one medium sized MEC (BigMec_e2) makes it infeasible.

The model hm_30 from [22] deserves special attention: It was handcrafted as an adversarial example for VI. Moreover, since the solvers for MDPs and MCs in PRISM use variants of VI, and since SI relies on those solvers, the PRISM implementation of SI also fails on the model¹⁰. QP shines on this model, being the only method to solve it. However, for increasing parameter N , the probabilities in hm_30 become so small that they are at the border of numerical stability. If the state-chains in the model were prolonged by one more state (i.e. the parameter N is set to 31), QP has numerical problems and reports an incorrect result. Similarly, noting that THOP reports an incorrect result on hm_30, one can experimentally find out that THOP succeeds only for $N \leq 25$. So if the model exhibits very small probabilities, one has to consider using a solver capable of arbitrary-precision arithmetic.

6 Conclusions

We have extended the three known classes of algorithm – value iteration, strategy iteration and quadratic programming – with several improvements and compared them both theoretically and practically.

In summary, for all algorithms, the structure of the underlying graph is more important than its size; thus knowledge about the model is relevant both for estimating the expected time, as well as the preferred algorithm and combination of optimizations. BVI and SI perform very similar on most models in our practical evaluation; each of them has some models where they are better. Quadratic/higher order programming is volatile and typically slower than the other two; however, the used solver has a huge impact, as we already see when changing between CPLEX, Gurobi and Minos. Thus, advances in the area of optimization problems could make this solution method the most practical.

A direction for future work is to extend all algorithms to other objectives, e.g. total expected reward, mean payoff or parity. Further, coming up with a polynomially sized convex QP would result in a polynomial-time algorithm, solving the long-standing open question.

¹⁰See [33, Appendix C.3] for more details on this.

References

- [1] Pranav Ashok, Krishnendu Chatterjee, Jan Kretínský, Maximilian Weininger & Tobias Winkler (2020): *Approximating Values of Generalized-Reachability Stochastic Games*. In: *LICS*, ACM, pp. 102–115, doi:10.1145/3373718.3394761.
- [2] Pranav Ashok, Jan Kretínský & Maximilian Weininger (2019): *PAC Statistical Model Checking for Markov Decision Processes and Stochastic Games*. In: *CAV (1), Lecture Notes in Computer Science 11561*, Springer, pp. 497–519, doi:10.1007/978-3-030-25540-4_29.
- [3] Christel Baier & Joost-Pieter Katoen (2008): *Principles of Model Checking*. MIT Press.
- [4] Christel Baier, Joachim Klein, Linda Leuschner, David Parker & Sascha Wunderlich (2017): *Ensuring the Reliability of Your Model Checker: Interval Iteration for Markov Decision Processes*. In: *CAV (1), Lecture Notes in Computer Science 10426*, Springer, pp. 160–180, doi:10.1007/978-3-319-63387-9_8.
- [5] Tomáš Brázdil, Krishnendu Chatterjee, Martin Chmelik, Vojtech Forejt, Jan Kretínský, Marta Z. Kwiatkowska, David Parker & Mateusz Ujma (2014): *Verification of Markov Decision Processes Using Learning Algorithms*. In: *ATVA, Lecture Notes in Computer Science 8837*, Springer, pp. 98–114, doi:10.1007/978-3-319-11936-6_8.
- [6] Krishnendu Chatterjee, Luca de Alfaro & Thomas A. Henzinger (2013): *Strategy improvement for concurrent reachability and turn-based stochastic safety games*. *J. Comput. Syst. Sci.* 79(5), pp. 640–657, doi:10.1016/j.jcss.2012.12.001.
- [7] Krishnendu Chatterjee & Nathanaël Fijalkow (2011): *A reduction from parity games to simple stochastic games*. In: *GandALF*, pp. 74–86, doi:10.4204/EPTCS.54.6.
- [8] Krishnendu Chatterjee, Kristoffer Arnsfelt Hansen & Rasmus Ibsen-Jensen (2017): *Strategy Complexity of Concurrent Safety Games*. In: *MFCS, LIPIcs 83*, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, pp. 55:1–55:13, doi:10.4230/LIPIcs.MFCS.2017.55.
- [9] Krishnendu Chatterjee & Thomas A. Henzinger (2008): *Value Iteration*. In: *25 Years of Model Checking, Lecture Notes in Computer Science 5000*, Springer, pp. 107–138, doi:10.1007/978-3-540-69850-0_7.
- [10] Krishnendu Chatterjee & Thomas A. Henzinger (2012): *A survey of stochastic ω -regular games*. *J. Comput. Syst. Sci.* 78(2), pp. 394–413, doi:10.1016/j.jcss.2011.05.002.
- [11] Krishnendu Chatterjee, Thomas A. Henzinger, Barbara Jobstmann & Arjun Radhakrishna (2010): *Gist: A Solver for Probabilistic Games*. In: *CAV, Lecture Notes in Computer Science 6174*, Springer, pp. 665–669, doi:10.1007/978-3-642-14295-6_57.
- [12] Krishnendu Chatterjee, Koushik Sen & Thomas A. Henzinger (2008): *Model-Checking omega-Regular Properties of Interval Markov Chains*. In: *FoSSaCS, Lecture Notes in Computer Science 4962*, Springer, pp. 302–317, doi:10.1007/978-3-540-78499-9_22.
- [13] Taolue Chen, Vojtech Forejt, Marta Z. Kwiatkowska, Aistis Simaitis & Clemens Wiltsche (2013): *On Stochastic Games with Multiple Objectives*. In: *MFCS, Lecture Notes in Computer Science 8087*, Springer, pp. 266–277, doi:10.1007/978-3-642-40313-2_25.
- [14] Chih-Hong Cheng, Alois Knoll, Michael Luttenberger & Christian Buckl (2011): *GAVS+: An Open Platform for the Research of Algorithmic Game Solving*. In: *TACAS, Lecture Notes in Computer Science 6605*, Springer, pp. 258–261, doi:10.1007/978-3-642-19835-9_22.
- [15] Anne Condon (1990): *On Algorithms for Simple Stochastic Games*. In: *Advances In Computational Complexity Theory, DIMACS Series in Discrete Mathematics and Theoretical Computer Science 13*, DIMACS/AMS, pp. 51–71, doi:10.1090/dimacs/013/04.
- [16] Anne Condon (1992): *The Complexity of Stochastic Games*. *Inf. Comput.* 96(2), pp. 203–224, doi:10.1016/0890-5401(92)90048-K.
- [17] Costas Courcoubetis & Mihalis Yannakakis (1995): *The Complexity of Probabilistic Verification*. *J. ACM* 42(4), pp. 857–907, doi:10.1145/210332.210339.

- [18] Decheng Dai & Rong Ge (2011): *Another Sub-exponential Algorithm for the Simple Stochastic Game*. *Algorithmica* 61(4), pp. 1092–1104, doi:10.1007/s00453-010-9413-1.
- [19] Peng Dai, Mausam, Daniel S. Weld & Judy Goldsmith (2011): *Topological Value Iteration Algorithms*. *J. Artif. Intell. Res.* 42, pp. 181–209. Available at <http://jair.org/papers/paper3390.html>.
- [20] J. Filar & K. Vrieze (1997): *Competitive Markov Decision Processes*. Springer-Verlag.
- [21] Hugo Gimbert & Florian Horn (2008): *Simple Stochastic Games with Few Random Vertices Are Easy to Solve*. In: *FoSSaCS, Lecture Notes in Computer Science* 4962, Springer, pp. 5–19, doi:10.1007/978-3-540-78499-9_2.
- [22] Serge Haddad & Benjamin Monmege (2018): *Interval iteration algorithm for MDPs and IMDPs*. *Theor. Comput. Sci.* 735, pp. 111–131, doi:10.1016/j.tcs.2016.12.003.
- [23] Ernst Moritz Hahn, Arnd Hartmanns, Christian Hensel, Michaela Klauck, Joachim Klein, Jan Křetínský, David Parker, Tim Quatmann, Enno Ruijters & Marcel Steinmetz (2019): *The 2019 Comparison of Tools for the Analysis of Quantitative Formal Models - (QComp 2019 Competition Report)*. In: *TACAS (3), Lecture Notes in Computer Science* 11429, Springer, pp. 69–92, doi:10.1007/978-3-030-17502-3_5.
- [24] Kristoffer Arnsfelt Hansen, Rasmus Ibsen-Jensen & Peter Bro Miltersen (2014): *The Complexity of Solving Reachability Games Using Value and Strategy Iteration*. *Theory Comput. Syst.* 55(2), pp. 380–403, doi:10.1007/s00224-013-9524-6.
- [25] Arnd Hartmanns & Benjamin Lucien Kaminski (2020): *Optimistic Value Iteration*. In: *CAV (2), Lecture Notes in Computer Science* 12225, Springer, pp. 488–511, doi:10.1007/978-3-030-53291-8_26.
- [26] A. J. Hoffman & R. M. Karp (1966): *On Nonterminating Stochastic Games*. *Management Science* 12(5), pp. 359–370, doi:10.1287/mnsc.12.5.359.
- [27] Rasmus Ibsen-Jensen & Peter Bro Miltersen (2012): *Solving Simple Stochastic Games with Few Coin Toss Positions*. In: *ESA, Lecture Notes in Computer Science* 7501, Springer, pp. 636–647, doi:10.1007/978-3-642-33090-2_55.
- [28] Mark Kattenbelt, Marta Z. Kwiatkowska, Gethin Norman & David Parker (2010): *A game-based abstraction-refinement framework for Markov decision processes*. *Formal Methods in System Design* 36(3), pp. 246–280, doi:10.1007/s10703-010-0097-6.
- [29] Edon Kelmendi, Julia Krämer, Jan Křetínský & Maximilian Weininger (2018): *Value Iteration for Simple Stochastic Games: Stopping Criterion and Learning Algorithm*. In: *CAV (1), Lecture Notes in Computer Science* 10981, Springer, pp. 623–642, doi:10.1007/978-3-319-96145-3_36.
- [30] Mikhail K Kozlov, Sergei P Tarasov & Leonid G Khachiyan (1980): *The polynomial solvability of convex quadratic programming*. *USSR Computational Mathematics and Mathematical Physics* 20(5), pp. 223–228, doi:10.1016/0041-5553(80)90098-1.
- [31] Jan Křetínský & Tobias Meggendorfer (2017): *Efficient Strategy Iteration for Mean Payoff in Markov Decision Processes*. In: *ATVA, Lecture Notes in Computer Science* 10482, Springer, pp. 380–399, doi:10.1007/978-3-319-68167-2_25.
- [32] Jan Křetínský & Tobias Meggendorfer (2019): *Of Cores: A Partial-Exploration Framework for Markov Decision Processes*. In: *CONCUR, LIPIcs* 140, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, pp. 5:1–5:17, doi:10.4230/LIPIcs.CONCUR.2019.5.
- [33] Jan Křetínský, Emanuel Ramneantu, Alexander Slivinskiy & Maximilian Weininger (2020): *Comparison of Algorithms for Simple Stochastic Games (Full Version)*. *CoRR* abs/2008.09465.
- [34] Marta Kwiatkowska, Gethin Norman, David Parker & Gabriel Santos (2020): *PRISM-games 3.0: Stochastic Game Verification with Concurrency, Equilibria and Time*. In: *CAV (2), Lecture Notes in Computer Science* 12225, Springer, pp. 475–487, doi:10.1007/978-3-030-53291-8_25.
- [35] Walter Ludwig (1995): *A Subexponential Randomized Algorithm for the Simple Stochastic Game Problem*. *Inf. Comput.* 117(1), pp. 151–155, doi:10.1006/inco.1995.1035.

- [36] Martin L. Puterman (1994): *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics, Wiley, doi:10.1002/9780470316887.
- [37] Tim Quatmann & Joost-Pieter Katoen (2018): *Sound Value Iteration*. In: *CAV (1), Lecture Notes in Computer Science* 10981, Springer, pp. 643–661, doi:10.1007/978-3-319-96145-3_37.
- [38] Rafal Somla (2005): *New Algorithms for Solving Simple Stochastic Games*. *Electron. Notes Theor. Comput. Sci.* 119(1), pp. 51–65, doi:10.1016/j.entcs.2004.07.008.
- [39] Marıa Svorenova & Marta Kwiatkowska (2016): *Quantitative verification and strategy synthesis for stochastic games*. *Eur. J. Control* 30, pp. 15–30, doi:10.1016/j.ejcon.2016.04.009.
- [40] Maximilian Weininger, Tobias Meggendorfer & Jan Kretınsky (2019): *Satisfiability Bounds for ω -Regular Properties in Bounded-Parameter Markov Decision Processes*. In: *CDC, IEEE*, pp. 2284–2291, doi:10.1109/CDC40024.2019.9029460.
- [41] Uri Zwick & Mike Paterson (1996): *The Complexity of Mean Payoff Games on Graphs*. *Theor. Comput. Sci.* 158(1&2), pp. 343–359, doi:10.1016/0304-3975(95)00188-3.