Approximating the Termination Value of One-Counter MDPs and Stochastic Games

Tomáš Brázdil[1]⋆, Václav Brožek[2]⋆⋆, Kousha Etessami[2], and Antonín Kučera[1]⋆

[1] Faculty of Informatics, Masaryk University
{xbrazdil,tony}@fi.muni.cz
[2] School of Informatics, University of Edinburgh
{vbrozek,kousha}@inf.ed.ac.uk

Abstract. One-counter MDPs (OC-MDPs) and one-counter simple stochastic games (OC-SSGs) are 1-player, and 2-player turn-based zero-sum, stochastic games played on the transition graph of classic one-counter automata (equivalently, pushdown automata with a 1-letter stack alphabet). A key objective for the analysis and verification of these games is the *termination* objective, where the players aim to maximize (minimize, respectively) the probability of hitting counter value 0, starting at a given control state and given counter value.

Recently [4, 2], we studied *qualitative* decision problems ("is the optimal termination value = 1?") for OC-MDPs (and OC-SSGs) and showed them to be decidable in P-time (in NP∩coNP, respectively). However, *quantitative* decision and approximation problems ("is the optimal termination value $\geq p$", or "approximate the termination value within ε") are far more challenging. This is so in part because optimal strategies may not exist, and because even when they do exist they can have a highly non-trivial structure. It thus remained open even whether any of these quantitative termination problems are computable.

In this paper we show that all quantitative *approximation* problems for the termination value for OC-MDPs and OC-SSGs are computable. Specifically, given a OC-SSG, and given $\varepsilon > 0$, we can compute a value v that approximates the value of the OC-SSG termination game within additive error ε, and furthermore we can compute ε-optimal strategies for both players in the game.

A key ingredient in our proofs is a subtle martingale, derived from solving certain LPs that we can associate with a maximizing OC-MDP. An application of Azuma's inequality on these martingales yields a computable bound for the "wealth" at which a "rich person's strategy" becomes ε-optimal for OC-MDPs.

1 Introduction

In recent years, there has been substantial research done to understand the computational complexity of analysis and verification problems for classes of finitely-presented but infinite-state stochastic models, MDPs, and stochastic games, whose transition

⋆ Tomáš Brázdil and Antonín Kučera are supported by the Institute for Theoretical Computer Science (ITI), project No. 1M0545, and by the Czech Science Foundation, grant No. P202/10/1469.

⋆⋆ Václav Brožek is supported by the Newton International Fellowship of Royal Society.

graphs arise from basic infinite-state automata-theoretic models, including: context-free processes, one-counter processes, and pushdown processes. It turns out these models are intimately related to important stochastic processes studied extensively in applied probability theory. In particular, one-counter probabilistic automata are basically equivalent to (discrete-time) quasi-birth-death processes (QBDs) (see [8]), which are heavily studied in queuing theory and performance evaluation as a basic model of an unbounded queue with multiple states (phases). It is very natural to extend these purely probabilistic models to MDPs and games, to model adversarial queuing scenarios.

In this paper we continue this work by studying quantitative *approximation* problems for **one-counter MDPs (OC-MDPs)** and **one-counter simple stochastic games (OC-SSGs)**, which are 1-player, and turn-based zero-sum 2-player, stochastic games on transition graphs of classic one-counter automata. In more detail, an OC-SSG has a finite set of control states, which are partitioned into three types: a set of *random* states, from where the next transition is chosen according to a given probability distribution, and states belonging to one of two players: *Max* or *Min*, from where the respective player chooses the next transition. Transitions can change the state and can also change the value of the (unbounded) counter by at most 1. If there are no control states belonging to *Max* (*Min*, respectively), then we call the resulting 1-player OC-SSG a *minimizing* (*maximizing*, respectively) OC-MDP. Fixing strategies for the two players yields a countable state Markov chain and thus a probability space of infinite runs (trajectories).

A central objective for the analysis and verification of OC-SSGs, is the *termination* objective: starting at a given control state and a given counter value $j > 0$, player Max (Min) wishes to maximize (minimize) the probability of eventually hitting the counter value 0 (in any control state). From well know fact, it follows that these games are *determined*, meaning they have a *value*, v, such that for every $\varepsilon > 0$, player Max (Min) has a strategy that ensures the objective is satisfied with probability at least $v - \varepsilon$ (at most $v + \varepsilon$, respectively), regardless of what the other player does. This value can be *irrational* even when the input data contains only rational probabilities, and this is so even in the purely stochastic case of QBDs without players ([8]).

A special subclass of OC-MDPs, called *solvency games*, was studied in [1] as a simple model of risk-averse investment. Solvency games correspond to OC-MDPs where there is only one control state, but there are multiple actions that change the counter value (“wealth”), possibly by more than 1 per transition, according to a finite support probability distribution on the integers associated with each action. The goal is to minimize the probability of going bankrupt, starting with a given positive wealth. It is not hard to see that these are subsumed by minimizing OC-MDPs (see [4]). It was shown in [1] that if the solvency game satisfies a number of restrictive assumptions (in particular, on the eigenvalues of a matrix associated with the game), then an optimal “rich person’s” strategy (which does the same action whenever the wealth is large enough) can be computed for it (in exponential time). They showed such strategies are not optimal for unrestricted solvency games and left the unrestricted case unresolved in [1].

We can classify analysis problems for OC-MDPs and OC-SSGs into two kinds. *Quantitative* analyses, which include: “is the game value at least/at most p” for a given $p \in [0, 1]$; or “approximate the game value” to within a desired additive error $\varepsilon > 0$. We can also restrict ourselves to *qualitative* analyses, which asks "is the game value = 1? = 0?".[3] We are also interested in strategies (e.g., memoryless, etc.) that achieve these.

In recent work [4, 2], we have studied *qualitative* termination problems for OC-SSGs. For both *maximizing* and *minimizing* OC-MDPs, we showed that these problems are decidable in P-time, using linear programming, connections to the theory of random walks on integers, and other MDP objectives. For OC-SSGs, we showed the qualitative termination problem "is the termination value = 1?" is in NP ∩ coNP. This problem is already as hard as Condon's quantitative termination problem for finite-state SSGs.

However we left open, as the main open question, the computability of *quantitative* termination problems for OC-MDPs and OC-SSGs. In this paper, we resolve positively the computability of all quantitative *approximation* problems associated with OC-MDPs and OC-SSGs. Note that, in some sense, approximation of the termination value in the setting of OC-MDPs and OC-SSGs can not be avoided. This is so not only because the value can be irrational, but because (see [3]) for maximizing OC-MDPs there need not exist any optimal strategy for maximizing the termination probability, only ε-optimal ones (whereas Min does have an optimal strategy in OC-SSGs). Moreover, even for minimizing OC-MDPs, where optimal strategies do exist, they can have a very complicated structure. In particular, as already mentioned for solvency games, there need not exist any "rich person's" strategy that can ignore the counter value when it is larger than some finite $N \geq 0$.

Nevertheless, we show all these difficulties can be overcome when the goal is to *approximate* the termination value of OC-SSGs and to compute ε-optimal strategies. Our main theorem is the following:

Theorem 1 **(ε-approximation of OC-SSG termination value).** *Given as input: a OC-SSG, $\mathcal{G}$, an initial control state s, an initial counter value $j > 0$, and a (rational) approximation threshold $\varepsilon > 0$, there is an algorithm that computes a rational number, v', such that $|v' - v^*| < \varepsilon$, where v^* is the value of the OC-SSG termination game on G, starting in configuration (s, j). Moreover, there is an algorithm that computes ε-optimal strategies for both players in the OC-SSG termination game. These algorithms run in exponential time in the encoding size of a 1-player OC-SSG, i.e., a OC-MDP, and in polynomial time in* $\log(1/\varepsilon)$ *and* $\log(j)$. *In the case of 2-player OC-SSGs, the algorithms run in nondeterministic exponential time in the encoding size of the OC-SSG.*

We now outline our basic strategy for proving this theorem. Consider the case of maximizing OC-MDPs, and suppose we would like to approximate the optimal termination probability, starting at state q and counter value i. Intuitively, it is not hard to see that as the counter value goes to infinity, except for some basic cases that we can detect and eliminate in polynomial time, the optimal probability of termination starting at a state q begins to approach the optimal probability of forcing the counter to have a $\liminf$ value $= -\infty$. But we can compute this optimal value, and an optimal strategy for it, based on results in our prior work [4, 2]. Of particular importance are the set of states T from which this value is 1. For a given $\varepsilon > 0$, we need to compute a bound N on the counter value, such that for any state q, and all counter values $N' > N$, the

[3] The problem "is the termination value = 0?" is easier, and can be solved in polynomial time without even looking at the probabilities labeling the transitions of the OC-SSG.

optimal termination probability starting at (q, N') is at most ε away from the optimal probability for the counter to have $\liminf$ value $= -\infty$. *A priori* it is not at all clear whether such a bound N is computable, although it is clear that N exists. To show that it is computable, we employ a subtle (sub)martingale, derived from solving a certain linear programming problem associated with a given OC-MDP. By applying Azuma's inequality on this martingale, we are able to show there are computable values $c < 1$, and $h \geq 0$, such that for all $i > h$, starting from a state q and counter value i, the optimal probability of both terminating and not encountering any state from which with probability 1 the player can force the $\liminf$ counter value to go to $-\infty$, is at most $c^i/(1-c)$. Thus, the optimal termination probability approaches from above the optimal probability of forcing the $\liminf$ counter value to be $-\infty$, and the difference between these two values is exponentially small in i, with a computable base c. This martingale argument extends to OC-MDPs an argument recently used in [7] for analyzing purely probabilistic one-counter automata (i.e., QBDs).

These bounds allow us to reduce the problem of approximating the termination value to the reachability problem for an exponentially larger finite-state MDP, which we can solve (in exponential time) using linear programming. The case for general OC-SSGs and minimizing OC-MDPs turns out to follow a similar line of argument, reducing the essential problem to the case of maximizing OC-MDPs. In terms of complexity, the OC-SSG case requires "guessing" an appropriate (albeit, exponential-sized) strategy, whereas the relevant exponential-sized strategy can be computed in deterministic exponential time for OC-MDPs. So our approximation algorithms run in exponential time for OC-MDPs and nondeterministic exponential time for OC-SSGs.

Open problems. An obvious remaining open problem is to obtain better complexity bounds for OC-MDPs. We know of no non-trivial lower bounds for OC-MDP approximation problems. Our results also leave open the decidability of the quantitative termination *decision* problem for OC-MDPs and OC-SSGs, which asks: "is the termination value $\geq p$?" for a given rational probability p. Furthermore, our results leave open computability for approximating the value of *selective termination* objectives for OC-MDPs, where the goal is to terminate (reach counter value 0) in a specific subset of the control states. Qualitative versions of selective termination problems were studied in [4, 2].

Related work. As noted, one-counter automata with a non-negative counter are equivalent to pushdown automata restricted to a 1-letter stack alphabet (see [8]), and thus OC-SSGs with the termination objective form a subclass of pushdown stochastic games, or equivalently, Recursive simple stochastic games (RSSGs). These more general stochastic games were studied in [9], where it was shown that many interesting computational problems, including any nontrivial approximation of the termination value for general RSSGs and RMDPs is undecidable, as are qualitative termination problems. It was also shown in [9] that for stochastic context-free games (1-exit RSSGs), which correspond to pushdown stochastic games with only one state, both qualitative and quantitative termination problems are decidable, and in fact qualitative termination problems are decidable in NP∩coNP ([10]), while quantitative termination problems are decidable in PSPACE. Solving termination objectives is a key ingredient for many more general analyses and model checking problems for such stochastic games (see, e.g., [5, 6]). OC-

SSGs are incompatible with stochastic context-free games. Specifically, for OC-SSGs, the number of stack symbols is bounded by 1, instead of the number of control states.

MDP variants of QBDs, essentially equivalent to OC-MDPs, have been considered in the queueing theory and stochastic modeling literature, see [14, 12]. However, in order to keep their analyses tractable, these works perform a naive finite-state "approximation" by cutting off the value of the counter at an arbitrary finite value N, and adding *dead-end absorbing* states for counter values higher than N. Doing this can radically alter the behavior of the model, even for purely probabilistic QBDs, and these authors establish no rigorous approximation bounds for their models. In a sense, our work can be seen as a much more careful and rigorous approach to finite approximation, employing at the boundary other objectives like maximizing the probability that the $\liminf$ counter value $= -\infty$. Unlike the prior work we establish rigorous bounds on how well our finite-state model approximates the original infinite OC-MDP.

2 Definitions

We assume familiarity with basic notions from probability theory. We call a probability distribution f over a discrete set, A, *positive* if $f(a) > 0$ for all $a \in A$, and *Dirac* if $f(a) = 1$ for some $a \in A$.

Definition 1 (SSG). *A* simple stochastic game (SSG) *is a tuple* $\mathcal{G} = (S, (S_0, S_1, S_2), \rightsquigarrow, Prob)$*, consisting of a countable set of* states*,* S*, partitioned into the set* S_0 *of* stochastic *states, and sets* S_1*,* S_2 *of states owned by Player 1 (Max) and 2 (Min), respectively. The* edge relation $\rightsquigarrow \subseteq S \times S$ *is total, i.e., for every* $r \in S$ *there is* $s \in S$ *such that* $r \rightsquigarrow s$*. Finally,* $Prob$ *assigns to every* $s \in S_0$ *a positive probability distribution over outgoing edges. If* $S_2 = \emptyset$*, we call the SSG a* maximizing *Markov Decision Processes (MDP). If* $S_1 = \emptyset$ *we call it a* minimizing *MDP.*

A *finite path* is a sequence $w = s_0 s_1 \cdots s_n$ of states such that $s_i \rightsquigarrow s_{i+1}$ for all $i, 0 \leq i < n$. We write $len(w) = n$ for the length of the path. A *run*, ω, is an infinite sequence of states every finite prefix of which is a path. For a finite path, w, we denote by $Run(w)$ the set of runs having w as a prefix. These generate the standard σ-algebra on the set of runs.

Definition 2 (OC-SSG). *A* one-counter SSG (OC-SSG)*,* $\mathcal{A} = (Q, (Q_0, Q_1, Q_2), \delta, P)$*, consists of a finite non-empty set of* control states*,* Q*, partitioned into stochastic and players' states, as in the case of SSGs, a set of* transition rules $\delta \subseteq Q \times \{+1, 0, -1\} \times Q$ *such that* $\delta(q) := \{(q, i, r) \in \delta\} \neq \emptyset$ *for all* $q \in Q$*, and* $P = \{P_q\}_{q \in Q_0}$ *where* P_q *is a positive rational probability distribution over* $\delta(q)$ *for all* $q \in Q_0$.

Purely for convenience, we assume that for each pair $q, r \in Q$ there is at most one i such that $(q, i, r) \in \delta$ (this is clearly w.l.o.g., by adding suitable auxiliary states to Q). By $\|\mathcal{A}\| := |Q| + |\delta| + \|P\|$ we denote the encoding size of $\mathcal{A}$, where $\|P\|$ is the sum of the number of bits needed to encode the numerator and denominator of $P_q(\varrho)$ for all $q \in Q$ and $\varrho \in \delta$. The set of all *configurations* is $C := \{(q, i) \mid q \in Q, i \geq 0\}$. Again, maximizing and minimizing OC-MDPs are defined as analogous subclasses of OC-SSGs.

To $\mathcal{A}$ we associate an infinite-state SSG $\mathcal{A}^\infty = (C, (C_0, C_1, C_2), \rightarrow, Prob)$, where the partition of C is defined by $(q, i) \in C_0$ iff $q \in Q_0$, and similarly for the players. The edges are defined by $(q, i) \rightarrow (r, j)$ iff either $i > 0$ and $(q, j - i, r) \in \delta$, or $i = j = 0$ and $q = r$. The probability assignment *Prob* is derived naturally from P.

By forgetting the counter values, the OC-SSG $\mathcal{A}$ also describes a finite-state SSG $\mathcal{G}_{\mathcal{A}} = (Q, (Q_0, Q_1, Q_2), \rightsquigarrow, Prob')$. Here $q \rightsquigarrow r$ iff $(q, i, r) \in \delta$ for some i, and $Prob'$ is derived in the obvious way from P by forgetting the counter changes. If $\mathcal{A}$ is a OC-MDP, both $\mathcal{G}_{\mathcal{A}}$ and $\mathcal{A}^{\infty}$ are MDPs.

Strategies and Probability. Let $\mathcal{G}$ be a SSG. A *history* is a finite path in $\mathcal{G}$. A *strategy* for Player 1 in $\mathcal{G}$, is a function assigning to each history ending in a state from S_1 a distribution on edges leaving the last state of the history. A strategy is *pure* if it always assigns a Dirac distribution, i.e., one which assigns 1 to one edge and 0 to the others. A strategy, σ, is *memoryless* if $\sigma(w) = \sigma(s)$ where s is the last state of a history w. Assume that $\mathcal{G} = \mathcal{A}^{\infty}$ for some OC-SSG $\mathcal{A}$. Then a strategy, σ, is *counterless* if it is memoryless and $\sigma((q, i)) = \sigma((q, 1))$ for all $i \geq 1$. Observe that every strategy, σ, for $\mathcal{G}_{\mathcal{A}}$ gives a unique strategy, σ', for $\mathcal{A}^{\infty}$; the strategy σ' just forgets the counter values in the history and plays as σ. This correspondence is bijective when restricted to memoryless strategies in $\mathcal{G}_{\mathcal{A}}$ and counterless strategies in $\mathcal{A}^{\infty}$. We will use this correspondence implicitly throughout the paper. Strategies for Player 2 are defined analogously.

Fixing a pair (σ, π) of strategies for Player 1 and 2, respectively, and an initial state, s, we obtain in a standard way a probability measure $\mathbb{P}^{\sigma,\pi}_{s}(\cdot)$ on the subspace of runs starting in s. For SSGs of the form $\mathcal{A}^{\infty}$ for some OC-SSG, $\mathcal{A}$, we consider two sequences of random variables, $\{C^{(i)}\}_{i \geq 0}$ and $\{S^{(i)}\}_{i \geq 0}$, returning the height of the counter, and the control state after completing i transitions.

For a SSG, $\mathcal{G}$, an *objective*, R, is for us a Borel subset of runs in $\mathcal{G}$. Player 1 is trying to *maximize* the probability of R, while player 2 is trying to *minimize* it. We say that $(\mathcal{G}, R)$ is *determined* if for every state s of $\mathcal{G}$ we have that $\sup_{\sigma} \inf_{\pi} \mathbb{P}^{\sigma,\pi}_{s}(R) = \inf_{\pi} \sup_{\sigma} \mathbb{P}^{\sigma,\pi}_{s}(R)$. If $(\mathcal{G}, R)$ is determined, then for every state s of $\mathcal{G}$, the above equality defines the *value* of s, denoted by $\mathrm{Val}(R, s)$. For a given $\varepsilon \geq 0$, a strategy, σ^*, of Player 1 is *ε-optimal* in s, if $\mathbb{P}^{\sigma^*,\pi}_{s}(R) \geq \mathrm{Val}(R, s) - \varepsilon$ for every strategy π of Player 2. An ε-optimal strategy for Player 2 is defined analogously. 0-optimal strategies are called *optimal*. Note that $(\mathcal{G}, R)$ is determined iff both players have ε-optimal strategies for every $\varepsilon > 0$.

Termination Objective. Let $\mathcal{A}$ be a OC-SSG. A run in $\mathcal{A}^{\infty}$ *terminates* if it contains a configuration of the form $(q, 0)$. The *termination objective* is the set of all terminating runs, and is denoted Term. OC-SSG termination games are determined (see [2]).

3 Main Result

Theorem 1 (Main). *Given an OC-SSG, $\mathcal{A}$, a configuration, (q, i), and a rational $\varepsilon > 0$, there is an algorithm that computes a rational number, v, such that $|\mathrm{Val}(\mathrm{Term}, (q, i)) - v| \leq \varepsilon$, and strategies σ, π for both players that are ε-optimal starting in (q, i). The algorithm runs in nondeterministic time exponential in $\|\mathcal{A}\|$ and polynomial in $\log(i)$ and $\log(1/\varepsilon)$. If $\mathcal{A}$ is an OC-MDP, then the algorithm runs in deterministic time exponential in $\|\mathcal{A}\|$ and polynomial in $\log(1/\varepsilon)$ and $\log(i)$.*

3.1 Proof sketch

We now sketch the main ideas in the proof of Theorem 1. First, observe that for all $q \in Q$ and $i \leq j$ we have that $\mathrm{Val}(\mathrm{Term}, (q,i)) \geq \mathrm{Val}(\mathrm{Term}, (q,j)) \geq 0$. Let

$$\mu_q := \lim_{i\to\infty} \mathrm{Val}(\mathrm{Term}, (q,i)).$$

Since $\mu_q \leq \mathrm{Val}(\mathrm{Term}, (q,i))$ for an arbitrarily large i, Player 1 should be able to decrease the counter by an arbitrary value with probability at least μ_q, no matter what Player 2 does. The objective of "decreasing the counter by an arbitrary value" cannot be formalized directly on $\mathcal{A}^\infty$, because the counter cannot become negative in the configurations of $\mathcal{A}^\infty$. Instead, we formalize this objective on $\mathcal{G}_{\mathcal{A}}$, extended with rewards on transitions. These rewards are precisely the counter changes, which were left out from the transition graph, i.e., each transition $q \rightsquigarrow r$ generated by a rule (q,i,r) of $\mathcal{A}$ has reward i. This allows us to define a sequence, $\{R^{(i)}\}_{i\geq 0}$, of random variables for runs in $\mathcal{G}_{\mathcal{A}}$, where $R^{(i)}$ returns the sum total of rewards accumulated during the first i steps. Note that $R^{(i)}$ may be negative, unlike the r.v. $C^{(i)}$. The considered objective then corresponds to the event $\mathit{LimInf}(=-\infty)$ consisting of all runs w in $\mathcal{G}_{\mathcal{A}}$ such that $\liminf_{i\to\infty} R^{(i)}(w) = -\infty$. These games are determined. For every $q \in Q$, let

$$\nu_q := \mathrm{Val}(\mathit{LimInf}(=-\infty), q).$$

One intuitively expects that $\mu_q = \nu_q$, and we show that this is indeed the case. Further, it was shown in [4, 2] that ν_q is rational and computable in non-deterministic time polynomial in $\|\mathcal{A}\|$. Moreover, both players have optimal pure memoryless strategies (σ^*, π^*) in $\mathcal{G}_{\mathcal{A}}$, computable in non-deterministic polynomial time. For MDPs, both the value ν_q and the optimal strategies can be computed in deterministic time polynomial in $\|\mathcal{A}\|$.

Since $\mu_q = \nu_q$, there is a sufficiently large N such that $\mathrm{Val}(\mathrm{Term}, (q,i)) - \nu_q \leq \varepsilon$ for all $q \in Q$ and $i \geq N$. We show that an upper bound on N is computable, which is at most exponential in $\|\mathcal{A}\|$ and polynomial in $\log(1/\varepsilon)$, in Section 3.2. As we shall see, this part is highly non-trivial. For all configurations (q,i), where $i \geq N$, the value $\mathrm{Val}(\mathrm{Term}, (q,i))$ can be approximated by ν_q, and both players can use the optimal strategies (σ^*, π^*) for the $\mathit{LimInf}(=-\infty)$ objective (which are "translated" into the corresponding counterless strategies in $\mathcal{A}^\infty$; cf. Section 2). For the remaining configurations (q,i), where $i < N$, we consider a finite-state SSG obtained by restricting $\mathcal{A}^\infty$ to configurations with counter between 0 and N, extended by two fresh stochastic states s_0, s_1 with self-loops. All configurations of the form $(q,0)$ have only one outgoing edge leading to s_0, and all configurations of the form (q,N) can enter either s_0 with probability ν_q, or s_1 with probability $1-\nu_q$. In this finite-state game, we compute the values and optimal strategies for the reachability objective, where the set of target states is $\{s_0\}$. This can be done in non-deterministic time polynomial in the size of the game (i.e., exponential in $\|\mathcal{A}\|$). If $\mathcal{A}$ is an OC-MDP, then the values and optimal strategies can be computed in deterministic polynomial time in the size of the MDP (i.e., exponential in $\|\mathcal{A}\|$) by linear programming (this applies both to the "maximizing" and the "minimizing" OC-MDP). Thus, we obtain the required approximations of $\mathrm{Val}(\mathrm{Term}, (q,i))$ for $i < N$, and the associated ε-optimal strategies.

Technically, we first consider the simpler case when $\mathcal{A}$ is a "maximizing" OC-MDP (Section 3.2). The general case is then obtained simply by computing the optimal counterless strategy π^* for the $\mathit{LimInf}(=-\infty)$ objective in $\mathcal{G}_{\mathcal{A}}$, and "applying" this strategy to resolve the choices of Player 2 in $\mathcal{A}^\infty$ (again, note that π^* corresponds to a counterless strategy in $\mathcal{A}^\infty$). Thus, we obtain an OC-MDP $\mathcal{A}'$ and apply the result of Section 3.2.

3.2 Bounding counter value N for maximizing OC-MDPs

In this section we consider a maximizing OC-MDP $\mathcal{A} = (Q, (Q_0, Q_1), \delta, P)$. The *maximum termination value* is $\mathrm{Val}(\mathrm{Term}, (q, i)) = \sup_\sigma \mathbb{P}^\sigma_{(q,i)}(\mathrm{Term})$.

For a $q \in Q$ we set $v_q := \sup_\sigma \mathbb{P}^\sigma_q(\mathit{LimInf}(=-\infty))$. Given $\mathcal{A}$, and $\varepsilon > 0$, we show here how to obtain a computable (exponential) bound on a number N such that $\left|\mathrm{Val}(\mathrm{Term}, (q, i)) - v_q\right| < \varepsilon$ for all $i \geq N$. Thus, by the arguments described in the Section 3, once we have such a computable bound on N, we have an algorithm for approximating $\mathrm{Val}(\mathrm{Term}, (q, i))$. We denote by T the set of all states q with $v_q = 1$.

Fact 2 (cf. [4]). *The number v_q is the max. probability of reaching T from q in $\mathcal{G}_{\mathcal{A}}$:*

$$v_q = \sup_\sigma \mathbb{P}^\sigma_q(\mathit{reach}\ T) = \max_\sigma \mathbb{P}^\sigma_q(\mathit{reach}\ T)\,.$$

Claim. $\forall q \in Q : \forall i \geq 0 : v_q \leq \mathrm{Val}(\mathrm{Term}, (q, i)) \leq \sup_\sigma \mathbb{P}^\sigma_{(q,i)}(\mathrm{Term} \cap \text{not reach } T) + v_q$.

Proof. The first inequality is easy. By [4, Theorem 12], $\mathrm{Val}(\mathrm{Term}, (q, i)) = 1$ for all $q \in T$, $i \geq 0$, proving the second inequality. □

Lemma 1. *Given a maximizing OC-MDP, $\mathcal{A}$, one can compute a rational constant $c < 1$, and an integer $h \geq 0$ such that for all $i \geq h$ and $q \in Q$:* $\sup_\sigma \mathbb{P}^\sigma_{(q,i)}(\mathrm{Term} \cap \mathit{not\ reach}\ T) \leq \frac{c^i}{1-c}$.
Moreover, $c \in \exp(1/2^{||\mathcal{A}||^{O(1)}})$ and $h \in \exp(||\mathcal{A}||^{O(1)})$.

Observe that this allows us to compute the number N. It suffices to set $N := \max\{h, \lceil \log_c(\varepsilon \cdot (1-c)) \rceil\}$. Based on the bounds on c and h, this allows us to conclude that $N \in \exp(||\mathcal{A}||^{O(1)})$, see [3]. In the rest of this section we prove Lemma 1.

We start with two preprocessing steps. First, we make sure that $T = \emptyset$, resulting in $\mathrm{Val}(\mathrm{Term}, (q, i)) = \sup_\sigma \mathbb{P}^\sigma_{(q,i)}(\mathrm{Term} \cap \text{not reach } T)$. Second, we make sure that there are no "degenerate" states in the system which would enable a strategy to spend an unbounded time with a bounded positive counter value. Both these reductions will be carried out in deterministic polynomial time.

In more detail, the first reduction step takes $\mathcal{A}$ and outputs $\mathcal{A}'$ given by replacing T with a single fresh control state, q_D ("D" for "diverging"), equipped with a single outgoing rule $(q_D, +1, q_D)$. By results of [4], this can be done in polynomial time. Obviously, for $q \notin T$ the value $\sup_\sigma \mathbb{P}^\sigma_{(q,i)}(\mathrm{Term} \cap \text{not reach } T)$ is the same in both $\mathcal{A}^\infty$ and $\mathcal{A}'^\infty$. Thus we may assume that $T = \emptyset$ when proving Lemma 1.

In the second reduction step, the property we need to assure holds in $\mathcal{A}$ is best stated in terms of $\mathcal{G}_{\mathcal{A}}$ and the variables $R^{(i)}$. We need to guarantee that under every pure memoryless strategy, $\liminf_{i\to\infty} R^{(i)}/i$ is almost surely positive. [4]

[4] This value is sometimes called the mean payoff, see also [4].

$$
\begin{aligned}
& z_q \leq -x + k + z_r && \text{for all } q \in Q_1 \text{ and } (q,k,r) \in \delta, \\
& z_q \leq -x + \textstyle\sum_{(q,k,r)\in\delta} P_q((q,k,r)) \cdot (k + z_r) && \text{for all } q \in Q_0, \\
& x > 0.
\end{aligned}
$$

Fig. 1. The system $\mathcal{L}$ of linear inequalities over x and z_q, $q \in Q$.

For runs starting in a state q, we denote by V_q the random variable giving the first time when q is revisited, or ∞ if it is not revisited. Let us call a pure memoryless strategy, σ, for $\mathcal{G}_{\mathcal{A}}$ *idling* if there is a state, q, such that $\mathbb{P}^{\sigma}_{q}\big(V_q < \infty\big) = 1$ and $\mathbb{P}^{\sigma}_{q}\big(R^{(V_q)} = 0\big) = 1$. We want to modify the OC-MDP, so that idling is not possible, without influencing the termination value. A technique to achieve this was already developed in our previous work [4], where we used the term “decreasing” for non-idling strategies. There we gave a construction which preserves the property of optimal termination probability being $= 1$. We in fact can establish that that construction preserves the exact termination value. Because the idea is not new, we leave details to [3].

After performing both reduction steps, we can safely assume that $T = \emptyset$ and that there are no idling pure memoryless strategies. The next claim then follows from Lemma 10 in [4]:

Claim. Under the assumptions above, for every pure memoryless strategy, σ, for $\mathcal{G}_{\mathcal{A}}$, and every $q \in Q$ we have $\mathbb{P}^{\sigma}_{q}\big(\liminf_{i\to\infty} R^{(i)}/i > 0\big) = 1$.

We shall now introduce a linear system of inequalities, $\mathcal{L}$, which is closely related to a standard LP that one can associate with a finite-state MDP with rewards, in order to obtain its optimal mean payoff value. The solution to the system of inequalities $\mathcal{L}$ will allow us to define a (sub)martingale that is critical for our arguments. This is an extension, to OC-MDPs, of a method used in [7] for analysis of purely probabilistic one-counter machines. The variables of $\mathcal{L}$ are x and z_q, for all $q \in Q$. The linear inequalities are defined in Figure 1.

Lemma 2. *There is a non-negative rational solution $(\bar{x}, (\bar{z}_q)_{q\in Q}) \in \mathbb{Q}^{|Q|+1}$ to $\mathcal{L}$, such that $\bar{x} > 0$. (The binary encoding size of the solution is polynomial in $\|\mathcal{A}\|$.)*

Proof. We first prove that there is some non-negative solution to $\mathcal{L}$ with $\bar{x} > 0$. The bound on size then follows by standard facts about linear programming. To find a solution, we will use optimal values of the MDP under the objective of minimizing *discounted total reward.* For every discount factor, λ, $0 < \lambda < 1$, there is a pure memoryless strategy, σ_λ, for $\mathcal{G}_{\mathcal{A}}$ such that $e^{\lambda}_{q}(\tau) := \sum_{i\geq 0} \lambda^i \cdot \mathbb{E}^{\tau}_{q}\big[R^{(i+1)} - R^{(i)}\big]$ is minimized by setting $\tau := \sigma_\lambda$. We prove that there is some λ, such that setting $\bar{z}_q := e^{\lambda}_{q}(\sigma_\lambda)$ and

$$
\begin{aligned}
\bar{x} := \min\big(&\{k + e^{\lambda}_{r}(\sigma_\lambda) - e^{\lambda}_{q}(\sigma_\lambda) \mid q \in Q_1, (q,k,r) \in \delta\} \\
&\cup \{P_q((q,k,r)) \cdot \big(k + e^{\lambda}_{r}(\sigma_\lambda) - e^{\lambda}_{q}(\sigma_\lambda)\big) \mid q \in Q_0, (q,k,r) \in \delta\}\big)
\end{aligned}
$$

forms a non-negative solution to $\mathcal{L}$ with $\bar{x} > 0$.

Now we proceed in more detail. By standard results (e.g., [13]), for a fixed state, q, and a fixed discount, $\lambda < 1$, there is always a pure memoryless strategy, σ_q, minimizing $e_q^\lambda(\tau)$ in place of τ. As we already proved in the Claim above, due to our assumptions we have $\mathbb{P}_q^{\sigma_q}\big(\liminf_{i\to\infty} R^{(i)}/i > 0\big) = 1$. Thus $\sum_{i\geq 0} \cdot \mathbb{E}_q^\tau\big[R^{(i+1)} - R^{(i)}\big] = \infty$, and there is a $\lambda < 1$ such that $e_q^\lambda(\sigma_q) > 0$ for all $q \in Q$. Finally, observe that there is a single strategy, σ_λ, which can be used as σ_q for all q. This is a consequence of σ_q being optimal also in successors of q. Finally, $\bar{x} > 0$, because for all $q \in Q_0$

$$\begin{aligned}
e_q^\lambda(\sigma_\lambda) &= \sum_{i\geq 0} \lambda^i \cdot \mathbb{E}_q^{\sigma_\lambda}\Big[R^{(i+1)} - R^{(i)}\Big] \\
&= \sum_{(q,k,r)\in\delta} P_q((q,k,r)) \cdot \left(k + \lambda \cdot \sum_{i\geq 0} \lambda^i \cdot \mathbb{E}_r^{\sigma_\lambda}\Big[R^{(i+1)} - R^{(i)}\Big]\right) \\
&= \sum_{(q,k,r)\in\delta} P_q((q,k,r)) \cdot \Big(k + \lambda \cdot e_r^\lambda(\sigma_\lambda)\Big) \\
&< \sum_{(q,k,r)\in\delta} P_q((q,k,r)) \cdot \Big(k + e_r^\lambda(\sigma_\lambda)\Big),
\end{aligned}$$

the last inequality following from $e_r^\lambda(\sigma_\lambda) > 0$ for all $r \in Q$; and similarly for all $q \in Q_1$ and $(q,k,r) \in \delta$

$$\begin{aligned}
e_q^\lambda(\sigma_\lambda) &= \sum_{i\geq 0} \lambda^i \cdot \mathbb{E}_q^{\sigma_\lambda}\Big[R^{(i+1)} - R^{(i)}\Big] \\
&\leq k + \lambda \cdot \sum_{i\geq 0} \lambda^i \cdot \mathbb{E}_r^{\sigma_\lambda}\Big[R^{(i+1)} - R^{(i)}\Big] = k + \lambda \cdot e_r^\lambda(\sigma_\lambda) < k + e_r^\lambda(\sigma_\lambda). \qquad \square
\end{aligned}$$

Recall the random variables $\{C^{(i)}\}_{i\geq 0}$ and $\{S^{(i)}\}_{i\geq 0}$, returning the height of the counter, and the control state after completing i transitions. Given the solution $(\bar{x}, (\bar{z}_q)_{q\in Q}) \in \mathbb{Q}^{|Q|+1}$ from Lemma 2, we define a sequence of random variables $\{m^{(i)}\}_{i\geq 0}$ by setting

$$m^{(i)} := \begin{cases} C^{(i)} + \bar{z}_{S^{(i)}} - i \cdot \bar{x} & \text{if } C^{(j)} > 0 \text{ for all } j,\ 0 \leq j < i, \\ m^{(i-1)} & \text{otherwise.} \end{cases}$$

We shall now show that $m^{(i)}$ defines a submartingale. For relevant definitions of (sub)martingales see, e.g., [11].

Lemma 3. *Under an arbitrary strategy, τ, for $\mathcal{A}^\infty$, and with an arbitrary initial configuration (q,n), the process $\{m^{(i)}\}_{i\geq 0}$ is a submartingale.*

Proof. Consider a fixed path, u, of length $i \geq 0$. For all j, $0 \leq j \leq i$ the values $C^{(j)}(\omega)$ are the same for all $\omega \in Run(u)$. We denote these common values by $C^{(j)}(u)$, and similarly for $S^{(j)}(u)$ and $m^{(j)}(u)$. If $C^{(j)}(u) = 0$ for some $j \leq i$, then $m^{(i+1)}(\omega) = m^{(i)}(\omega)$ for every $\omega \in Run(u)$. Thus $\mathbb{E}_{(q,n)}^\tau\big[m^{(i+1)} \mid Run(u)\big] = m^{(i)}(u)$. Otherwise, consider the last configuration, (r,l), of u. For every possible successor, (r',l'), set

$$p_{(r',l')} := \begin{cases} \tau(u)((r,l) \to (r',l')) & \text{if } r \in Q_1, \\ Prob((r,l) \to (r',l')) & \text{if } r \in Q_0. \end{cases}$$

Then

$$\mathbb{E}^{\tau}_{(q,n)}\Big[C^{(i+1)} - C^{(i)} + \bar{z}_{S^{(i+1)}} - \bar{x} \mid Run(u)\Big] = \quad -\bar{x} + \sum_{(r,k,r')\in\delta} p_{(r',l+k)} \cdot (k + \bar{z}_{r'}) \quad \geq \quad \bar{z}_r.$$

This allows us to derive the following:

$$\begin{aligned}\mathbb{E}^{\tau}_{(q,n)}\Big[m^{(i+1)} \mid Run(u)\Big] &= \mathbb{E}^{\tau}_{(q,n)}\Big[C^{(i+1)} + \bar{z}_{S^{(i+1)}} - (i+1)\cdot\bar{x} \mid Run(u)\Big] \\ &= C^{(i)}(u) + \mathbb{E}^{\tau}_{(q,n)}\Big[C^{(i+1)} - C^{(i)} + \bar{z}_{S^{(i+1)}} - \bar{x} \mid Run(u)\Big] - i\cdot\bar{x} \\ &\geq C^{(i)}(u) + \bar{z}_{S^{(i)}(u)} - i\cdot\bar{x} \quad = \quad m^{(i)}(u).\end{aligned}$$ □

Now we can finally prove Lemma 1. Denote by Term_j the event of terminating after *exactly* j steps. Further set $\bar{z}_{\max} := \max_{q\in Q}\bar{z}_q - \min_{q\in Q}\bar{z}_q$, and assume that $C^{(0)} \geq \bar{z}_{\max}$. Then the event Term_j implies that $m^{(j)} - m^{(0)} = \bar{z}_{S^{(j)}} - j\cdot\bar{x} - C^{(0)} - \bar{z}_{S^{(0)}} \leq -j\cdot\bar{x}$. Finally, observe that we can bound the one-step change of the submartingale value by $\bar{z}_{\max} + \bar{x} + 1$. Using the Azuma-Hoeffding inequality for the submartingale $\{m^{(n)}\}_{n\geq 0}$ (see, e.g., Theorem 12.2.3 in [11]), we thus obtain the following bound for every strategy σ and initial configuration (q,i) with $i \geq \bar{z}_{\max}$:

$$\mathbb{P}^{\sigma}_{(q,i)}\Big(\mathrm{Term}_j\Big) \leq \mathbb{P}^{\sigma}_{(q,i)}\Big(m^{(j)} - m^{(0)} \leq -j\cdot\bar{x}\Big) \leq \exp\left(\frac{-\bar{x}^2\cdot j^2}{2j\cdot(\bar{z}_{\max}+\bar{x}+1)}\right).$$

We choose $c := \exp\left(\frac{-\bar{x}^2}{2\cdot(\bar{z}_{\max}+\bar{x}+1)}\right) < 1$ and observe that

$$\mathbb{P}^{\sigma}_{(q,i)}(\mathrm{Term}) = \sum_{j\geq i}\mathbb{P}^{\sigma}_{(q,i)}\Big(\mathrm{Term}_j\Big) \leq \sum_{j\geq i} c^j = \frac{c^i}{1-c}.$$

This choice of c, together with $h := \lceil\bar{z}_{\max}\rceil$, finishes the proof of Lemma 1. (The given bounds on c and h are easy to check.) □

3.3 Bounding *N* for general SSGs

For a control state q, let $v_q := \sup_\sigma \inf_\pi \mathbb{P}^{\sigma,\pi}_q(\mathit{LimInf}(=-\infty))$. Given a OC-SSG, $\mathcal{A} = (Q,(Q_0,Q_1,Q_2),\delta,P)$, and $\varepsilon > 0$, we now show how to obtain a computable bound on the number N such that $\left|\mathrm{Val}(\mathrm{Term},(q,i)) - v_q\right| < \varepsilon$ for all $i \geq N$. Again, by the arguments described in Section 3, once we have this, we have an algorithm for approximating $\mathrm{Val}(\mathrm{Term},(q,i))$.

By results in [2], there is always a counterless pure strategy, π^*, for Player 2 in $\mathcal{G}_{\mathcal{A}}$, such that

$$\sup_\sigma \inf_\pi \mathbb{P}^{\sigma,\pi}_q(\mathit{LimInf}(=-\infty)) = \sup_\sigma \mathbb{P}^{\sigma,\pi^*}_q(\mathit{LimInf}(=-\infty)).$$

Observe that by fixing the choices of π^* in $\mathcal{A}$ we obtain a maximizing OC-MDP, $\mathcal{A}^* = (Q^*,(Q^*_0,Q^*_1),\delta^*,P^*)$, where $Q^*_0 = Q_0 \cup Q_2$, $Q^*_1 = Q_1$, $\delta^* := \{(q,k,r)\in\delta \mid q \in Q_0 \cup Q_1 \vee \pi^*(q) = r\}$, and P^* is the unique (for $\mathcal{A}^*$) extension of P to states from Q_2.

Slightly abusing notation, denote also by π^* the strategy for $\mathcal{A}^\infty$ which corresponds, in the sense explained in Section 2, to π^* for $\mathcal{G}_{\mathcal{A}}$. Then $v_q = \mathbb{P}^{\sigma,\pi^*}_q(\mathit{LimInf}(=-\infty)) \leq \mathrm{Val}(\mathrm{Term},(q,i)) \leq \mathbb{P}^{\sigma,\pi^*}_{(q,i)}(\mathrm{Term})$. Applying Lemma 1 to the OC-MDP $\mathcal{A}^*$ thus allows us to give a computable (exponential) bound on N, given $\mathcal{A}$.

References

1. Berger, N., Kapur, N., Schulman, L.J., Vazirani, V.: Solvency Games. In: Proc. of FSTTCS'08 (2008)
2. Brázdil, T., Brožek, V., Etessami, K.: One-Counter Simple Stochastic Games. In: Proc. of FSTTCS'10. pp. 108–119 (2010)
3. Brázdil, T., Brožek, V., Etessami, K., Kučera, A.: Approximating the Termination Value of One-Counter MDPs and Stochastic Games. Tech. Rep. abs/1104.4978, CoRR, http://arxiv.org/abs/1104.4978 (2011)
4. Brázdil, T., Brožek, V., Etessami, K., Kučera, A., Wojtczak, D.: One-Counter Markov Decision Processes. In: ACM-SIAM SODA. pp. 863–874 (2010), full tech report: CoRR, abs/0904.2511, 2009. http://arxiv.org/abs/0904.2511.
5. Brázdil, T., Brožek, V., Forejt, V., Kučera, A.: Reachability in recursive Markov decision processes. In: Proc. 17th Int. CONCUR. pp. 358–374 (2006)
6. Brázdil, T., Brožek, V., Kučera, A., Obdržálek, J.: Qualitative Reachability in stochastic BPA games. In: Proc. 26th STACS. pp. 207–218 (2009)
7. Brázdil, T., Kiefer, S., Kučera, A.: Efficient analysis of probabilistic programs with an unbounded counter. CoRR abs/1102.2529 (2011)
8. Etessami, K., Wojtczak, D., Yannakakis, M.: Quasi-birth-death processes, tree-like QBDs, probabilistic 1-counter automata, and pushdown systems. In: Proc. 5th Int. Symp. on Quantitative Evaluation of Systems (QEST). pp. 243–253 (2008)
9. Etessami, K., Yannakakis, M.: Recursive Markov decision processes and recursive stochastic games. In: Proc. 32nd ICALP. pp. 891–903 (2005)
10. Etessami, K., Yannakakis, M.: Efficient qualitative analysis of classes of recursive Markov decision processes and simple stochastic games. In: Proc. of 23rd STACS'06. Springer (2006)
11. Grimmett, G.R., Stirzaker, D.R.: Probability and Random Processes. Oxford U. Press, 2nd edn. (1992)
12. Lambert, J., Van Houdt, B., Blondia, C.: A policy iteration algorithm for markov decision processes skip-free in one direction. In: ValueTools. ICST, Brussels, Belgium (2007)
13. Puterman, M.L.: Markov Decision Processes. J. Wiley and Sons (1994)
14. White, L.B.: A new policy iteration algorithm for Markov decision processes with quasi birth-death structure. Stochastic Models 21, 785–797 (2005)