Approximating the termination value of one-counter MDPs and stochastic games

Brazdil, T., Brozek, V., Etessami, K., & Kucera, A. (2013). Approximating the termination value of one-counter MDPs and stochastic games. Information and Computation, 222, 121-138. 10.1016/j.ic.2012.01.008

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

Tomáš Brázdil[1,1], Václav Brožek[b,1], Kousha Etessami[b], Antonín Kučera[1,1]

[a]*Faculty of Informatics, Masaryk University,*
Botanická 68a, 60200 Brno
Czech Republic
[b]*School of Informatics, University of Edinburgh*
Informatics Forum
10 Crichton Street
EH8 9AB, Edinburgh
United Kingdom

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, we studied *qualitative* decision problems ("is the optimal termination value equal to 1?") for OC-MDPs (and OC-SSGs) and showed them to be decidable in polynomial time (in NP ∩ coNP, respectively). However, *quantitative* decision and approximation problems ("is the optimal termination value at least 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 ter-

[1]Authors supported by the research center Institute for Theoretical Computer Science, project No. 1M0545. Tomáš Brázdil and Antonín Kučera are also supported by the Czech Science Foundation, project No. P202/10/1469. Václav Brožek is also supported by Newton International Fellowship from the Royal Society.

mination 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 linear programs 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 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 [9]), 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 ([9]).

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 [3]). 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?".[2] We are also interested in strategies (e.g., memoryless, etc.) that achieve these.

In recent work [2, 3], 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

[2]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.

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.

Our contribution. 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 Example A.1 in Section A.1) 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 result* (Theorem 3.1) is the following:

There is an algorithm that, given as input: a OC-SSG, $\mathcal{G}$, an initial control state s, an initial counter value $j > 0$, and a (rational) approximation threshold $\varepsilon > 0$,

- *computes a rational number, v', such that $|v' - v^*| < \varepsilon$, where v^* is the value of the OC-SSG termination game on $\mathcal{G}$, starting in configuration (s, j), and*
- *computes ε-optimal strategies for both players in the OC-SSG termination game.*

For OC-MDPs, i.e., 1-player OC-SSGs, the algorithm runs in exponential time in the encoding size of the OC-MDP, and in polynomial time in $\log(1/\varepsilon)$ and $\log(j)$. For 2-player OC-SSGs, the algorithm runs in nondeterministic exponential time in the encoding size of the OC-SSG.[3]

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

[3]We shall explain after the statement of Theorem 3.1, in footnote 4, p precisely what we mean by computing something in *nondeterministic* exponential time. It amounts to the standard notion of nondeterministic computation used in the setting of total search problems.

termination probability, starting at state q and counter value i. Intuitively, it is not hard to believe that as the counter value goes to infinity, the optimal probability of termination starting at a state q begins to approach the optimal probability, v_q, of forcing the counter to have a lim inf value $= -\infty$. We prove that this is indeed the case. But we can compute the optimal value v_q and an optimal strategy for achieving it, based on results in our prior work [2, 3]. 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 optimal termination probability starting at (q, N') is at most ε away from the optimal probability for the counter to have lim inf value $= -\infty$. *A priori* it is not 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 lim inf 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 lim inf 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 [6] 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.

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 [9]), 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 [10], 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 [10] 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 ([11]), 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., [4, 5]). 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, 17]. 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 lim inf 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. Preliminaries

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 $a \in A$.

Definition 2.1. A *One-Counter Simple Stochastic Game (OC-SSG)* is given as $\mathcal{A} = (Q, \Delta, P)$, where

- Q is a finite non-empty set of *control states*, partitioned into the states $Q_\top$ of player Max, $Q_\bot$ of player Min, and stochastic states Q_P;
- a set $\Delta \subseteq Q \times \{-1, 0, +1\} \times Q$ of *transition rules*, such that for all $q \in Q$ there is some $(q, a, r) \in \delta$;
- a map P taking each tuple $(q, a, r) \in \Delta$ with $q \in Q_P$ to a positive rational number $P((q, a, r))$, so that for every $q \in Q_P$: $\sum_{(q,a,r)\in\delta} P((q, a, r)) = 1$.

A *configuration* is a pair (q, c) of a control state, q, and an integer counter value $c \in \mathbb{Z}$. The set of all configurations is $Q \times \mathbb{Z}$. An OC-SSG where $Q_\perp = \emptyset$ is called a *maximizing One-Counter Markov Decision Process (maximizing OC-MDP)*, similarly $Q_\top = \emptyset$ defines a *minimizing OC-MDP*. Finally, if $Q_\top = Q_\perp = \emptyset$ we have a *One-Counter Markov Chain (OC-MC)*.

Let us fix a OC-SSG, $\mathcal{A} = (Q, \Delta, P)$. A *run* in $\mathcal{A}$ is an infinite sequence of configurations $\omega = (q_0, c_0)(q_1, c_1) \cdots$ such that for all $i \geq 1$ we have that $(q_{i-1}, c_i - c_{i-1}, q_i) \in \Delta$. We define for every $n \geq 0$ the following functions:

- $\mathrm{State}^{(n)} : \mathit{Run} \to Q$ returns the n-th control state: $\mathrm{State}^{(n)}(\omega) = q_n$.
- $\mathrm{C}^{(n)} : \mathit{Run} \to \mathbb{Z}$ returns the n-th counter value: $\mathrm{C}^{(n)}(\omega) = c_n$.

A finite prefix, $w = (q_0, c_0) \cdots (q_k, c_k)$, of a run is called a *finite path*, and $\mathit{len}(w) := k$ is its length. We denote by *Run* the set of all runs, and by $\mathit{Run}(w)$ the set of all runs starting with a finite path w. Closing the set $\{\mathit{Run}(w) \mid w \text{ is a finite path}\}$ under complements and countable unions generates the standard Borel σ-algebra of measurable sets of runs. Note that the functions $\mathrm{State}^{(n)}$ and $\mathrm{C}^{(n)}$ have measurable pre-images.

A *strategy* for player Max is a function, σ, which to each finite path $w = (q_0, c_0) \cdots (q_k, c_k)$, also called *history* in this context, where $q_k \in Q_\top$, assigns a probability distribution on the set of rules of the form $(q_k, a, r) \in \Delta$. It is called *pure* if $\sigma(w)$ assigns probability 1 to some transition, for each history w. We call σ *counterless* if $\sigma(w)$ depends only on the last control state, q_k. Strategies for Min are defined similarly, just by substituting $Q_\top$ with $Q_\perp$.

Assume that a pair (σ, π) of strategies for Max and Min, respectively, is fixed. Consider a finite path $w = (q_0, c_0) \cdots (q_k, c_k)$ and a rule $(q_{i-1}, c_i - c_{i-1}, q_i) \in \Delta$, $1 \leq i \leq k$. We assign to this rule a weight, x_i, as follows: If $q_{i-1} \in Q_P$ then $x_i = \mathit{Prob}((q_{i-1}, c_i - c_{i-1}, q_i))$. If $q_{i-1} \in Q_\top$ then x_i is equal to the probability of $(q_{i-1}, c_i - c_{i-1}, q_i)$ assigned by $\sigma((q_0, c_0) \cdots (q_{i-1}, c_{i-1}))$, and similarly for $q_{i-1} \in Q_\perp$ and π. The weight of w is then $x_w = \prod_{i=1}^{\mathit{len}(w)-1} x_i$, where the empty product is equal to 1. Once we also fix an *initial configuration*, (q, c), we obtain a probability measure $\mathbb{P}^{\sigma,\pi}_{(q,c)}$. This is defined by setting $\mathbb{P}^{\sigma,\pi}_{(q,c)}(\mathit{Run}(w)) = 0$ if w does not start with (q, c), and $\mathbb{P}^{\sigma,\pi}_{(q,c)}(\mathit{Run}(w)) = x_w$ if w starts with (q, c). This and the requirement of countable additivity of a measure already uniquely describes $\mathbb{P}^{\sigma,\pi}_{(q,c)}$ (see, e.g., [16, p. 30] for the case of MDPs. The extension of this to SSGs is straightforward.) If $\mathcal{A}$ is a maximizing OC-MDP, a minimizing OC-MDP, or a OC-MC, we denote the probability measure by $\mathbb{P}^{\sigma}_{(q,c)}$, $\mathbb{P}^{\pi}_{(q,c)}$, or $\mathbb{P}_{(q,c)}$, respectively.

Objectives. In this paper, an *objective* for an OC-SSG is a measurable set of runs. Player Max tries to maximize the probability of this set, whereas player Min tries to minimize it. Given an objective, O, for a OC-SSG, $\mathcal{A}$, and a configuration, $s = (q, c)$, we define the *value in s* as

$$\mathrm{Val}_{\mathcal{A}}(O, s) := \sup_{\sigma} \inf_{\pi} \mathbb{P}_s^{\sigma,\pi}(O) = \inf_{\pi} \sup_{\sigma} \mathbb{P}_s^{\sigma,\pi}(O)\,.$$

The latter equality follows from Martin's Blackwell determinacy theorem [15]. We write just $\mathrm{Val}(O, s)$ if $\mathcal{A}$ is understood. For an $\varepsilon \geq 0$, a strategy σ for Max is called *ε-optimal in s* if $\mathbb{P}_s^{\sigma,\pi}(O) \geq \mathrm{Val}(O, s) - \varepsilon$ for every π. Similarly a strategy π for Min is *ε-optimal in s* if $\mathbb{P}_s^{\sigma,\pi}(O) \leq \mathrm{Val}(O, s) + \varepsilon$ for every σ. A 0-optimal strategy is called *optimal*. Note that by determinacy both players have ε-optimal strategies for every $\varepsilon > 0$.

The key objective is the *termination* objective:

$$\mathit{Term} := \{\omega \in \mathit{Run} \mid \exists n : \mathrm{C}^{(n)}(\omega) \leq 0\}.$$

The name "termination" stems from the connection to one-counter automata. Such automata also have a finite number of control states and a non-negative counter, and a run can be considered to "terminate" upon hitting counter value 0. OC-SSGs do not necessarily halt when the counter is 0, and allow negative counter values. However, this difference is irrelevant from the perspective of the termination objective, for which only the part of runs with non-negative counter values matter.

Games without a Counter. In our arguments we also use the notion of Simple Stochastic Games (SSGs) of Condon [7], which are similar to OC-SSGs. The main difference is the lack of a counter, and the focus on the objective of reaching a distinguished sink state.

Definition 2.2. A *simple stochastic game (SSG)* is a tuple $\mathcal{G} = (S, \rightsquigarrow, \mathit{Prob})$, where

- S is a finite set of *states*, partitioned into the states $S_\top$ of player Max, $S_\bot$ of player Min, and stochastic states S_P;
- $\rightsquigarrow \subseteq S \times S$ is a transition relation such that for every state $s \in S$ there is at least one state $r \in S$ such that $s \rightsquigarrow r$;

- *Prob* is a *probability assignment* which to each $s \in S_P$ assigns a rational probability distribution on its set of successors, where for a state $s \in S_P$ its successors are defined to be the set $\{r \mid s \rightsquigarrow r\}$.

If $S_\perp = \emptyset$ we call $\mathcal{G}$ a *maximizing Markov decision process (maximizing MDP)*. If $S_\top = \emptyset$ we call it a *minimizing MDP*. If $S_\top = S_\perp = \emptyset$ we call $\mathcal{G}$ a *Markov chain*.

The SSG also comes with a distinguished sink state $s_0 \in S$, and this implicitly defines the *reachability objective* "reach s_0" defined by runs ω which visit s_0.

Runs, strategies, probability measures and values with respect to objectives are defined analogously to those for OC-SSGs, just by removing references to the counter. In particular, runs are sequences of states. The following is well known.

Fact 2.3. *(See, e.g., [7, 8, 16].) For both maximizing and minimizing MDPs, optimal pure memoryless strategies for reachability exist and can be computed, together with the optimal reachability value, in polynomial time.*

3. Main Result

Theorem 3.1 (Main)**.** *There is an algorithm that, given an OC-SSG, $\mathcal{A}$, a configuration, (q, i), $i \geq 0$, and a rational $\varepsilon > 0$, computes a rational number, ν, such that $|\mathrm{Val}(\mathit{Term}, (q, i)) - \nu| \leq \varepsilon$, and computes strategies σ and π for the Max and Min player, respectively, such that both σ and π are ε-optimal starting in (q, i) with respect to the termination objective. 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 $\log(i)$ and $\log(1/\varepsilon)$.*[4]

[4] To make precise the meaning of this theorem, we have to spell out precisely what we mean by a *nondeterministic* algorithm that computes ν, σ and τ within given resource bounds. This is a standard notion for total search problems. We will say a nondeterministic algorithm (i.e., nondeterministic Turing machine) computes ν, σ and τ within the specified resource bounds (namely, exponential time) if given any input the algorithms halts in exponential time on all computation paths, and furthermore, if the input is not well-formed the algorithm "rejects" it on all computation paths, but if the input is well-formed (i.e., if it is given a well formed OC-SSG, $\mathcal{A}$, initial configuration (q, i), and $\epsilon > 0$) the nondeterministic algorithm: (a) has at least one accepting computation path; and (b) on every accepting computation path it outputs values ν, σ, and π which satisfy that $|\mathrm{Val}(\mathit{Term}, (q, i)) - \nu| \leq \varepsilon$ and such that σ and π are ϵ-optimal strategies for Max and Min, respectively, for the given input OC-SSG $\mathcal{A}$ and initial configuration (q, i). On rejecting computation paths the algorithm need not output anything. Note that the outputs on different accepting executions may be different, but they must all satisfy the required specification.

Let us first briefly sketch the main ideas in the proof of Theorem 3.1. First, observe that for all $q \in Q$ and $i \leq j$ we have that $\mathrm{Val}(\mathit{Term}, (q, i)) \geq \mathrm{Val}(\mathit{Term}, (q, j)) \geq 0$. Let

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

Since $\mu_q \leq \mathrm{Val}(\mathit{Term}, (q, i))$ for an arbitrarily large i, Player Max should be able to decrease the counter by an arbitrary value with probability at least μ_q, no matter what Player Min does. The objective of "decreasing the counter by an arbitrary value" can be formalized directly as the following "limit" objective, which has useful connections to termination [3]:

$$\mathit{LimInf}(= -\infty) \coloneqq \{\omega \in \mathit{Run} \mid \liminf_{n \to \infty} \mathrm{C}^{(n)}(\omega) = -\infty\}.$$

OC-SSG with this objective are determined, which means that the following value is defined for every $q \in Q$:

$$\nu_q \coloneqq \mathrm{Val}(\mathit{LimInf}(= -\infty), (q, n)), \quad \text{where } n = 0. \tag{1}$$

Remark 3.1. Observe that due to the nature of $\mathit{LimInf}(= -\infty)$ we would obtain the same value ν_q if we used any other value of n. It will be often the case that we will measure the (optimal) probability of some events, where the resulting number will not depend on the initial counter value. From now on, in such cases we will specify only the initial state, so, e.g., (1) would become $\nu_q \coloneqq \mathrm{Val}(\mathit{LimInf}(= -\infty), q)$.

One intuitively expects that $\mu_q = \nu_q$, and we show that this is indeed the case (see Corollary 3.13). Further, by [2, Theorem 2], ν_q is rational and computable in non-deterministic time polynomial in $\|\mathcal{A}\|$. Moreover, both players have optimal pure counterless strategies (σ^*, π^*) computable in non-deterministic polynomial time. For OC-MDPs, both the value ν_q and the optimal strategies can be computed in deterministic time polynomial in $\|\mathcal{A}\|$.

Obviously, there must be a (sufficiently large) N such that $\mathrm{Val}(\mathit{Term}, (q, i)) - \mu_q \leq \varepsilon$ for all $q \in Q$ and $i \geq N$. We show that an upper bound on N is computable, and is at most exponential in $\|\mathcal{A}\|$ and polynomial in $\log(1/\varepsilon)$, in Section 3.1. As we shall see, this part is highly non-trivial. For all configurations (q, i), where $i \geq N$, the value $\mathrm{Val}(\mathit{Term}, (q, i))$ can be approximated by μ_q $(= \nu_q)$, and both players can use the optimal strategies (σ^*, π^*) for the $\mathit{LimInf}(= -\infty)$ objective. For the remaining configurations (q, i), where $i < N$, we consider a (finite-state) SSG $\mathcal{G}$ obtained by restricting ourselves 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 SSG, we compute the values and optimal strategies for the objective of reaching s_0. This can be done in nondeterministic time polynomial in the size of $\mathcal{G}$ (i.e., exponential in $\|\mathcal{A}\|$). If $\mathcal{A}$ is an OC-MDP, then $\mathcal{G}$ is a MDP, and the values and optimal strategies can be computed in deterministic polynomial time in the size of $\mathcal{G}$ (i.e., exponential in $\|\mathcal{A}\|$) by linear programming (this applies both to the "maximizing" and the "minimizing" OC-MDPs). Thus, we obtain the required approximations of $\mathrm{Val}(\mathit{Term}, (q, i))$ for $i < N$, and the associated ε-optimal strategies.

Proof of Theorem 3.1. The algorithm is given an OC-SSG $\mathcal{A} = (Q, \Delta, P)$, an initial configuration (q, i), and a rational number $\varepsilon > 0$, as input. Recall that for $r \in Q$ we set $\nu_r := \mathrm{Val}(\mathit{LimInf}(= -\infty), r)$. The algorithm does the following:

1. Compute a pair (σ^*, π^*) of pure counterless strategies, for players Max and Min, respectively, which are optimal for $\mathit{LimInf}(= -\infty)$ starting at every state $r \in Q$. Compute ν_r, for every $r \in Q$.
2. Compute N such that $\mathrm{Val}(\mathit{Term}, (r, j)) - \nu_r \leq \varepsilon$ for all $r \in Q$ and $j \geq N$.
3. If $i \geq N$ then return ν_q, σ^*, π^*.
4. Otherwise apply the algorithm from Lemma 3.14 to $\mathcal{A}$, ε, $(\nu_r)_{r \in Q}$, N, σ^*, π^* and return $\nu_{(q,i)}$, $\overline{\sigma}$, $\overline{\pi}$ from its output.

A key step is obviously step 4, which is not described here. We shall describe and prove the correctness and complexity of that step in Lemma 3.14. If we can carry out the computations as specified in Steps 1 and 2, then the correctness of the output in Step 3 holds by definition. Let us now evaluate the complexity of the first two steps (using some of our earlier results, and some results that will be established in Section 3.1):

- (Step 1.) The values ν_r, $r \in Q$, which have polynomially big encoding by [2, Proposition 9], can be guessed and verified in polynomial time by [2, Theorem 2]. Strategies exist that are optimal with respect to $\mathit{LimInf}(= -\infty)$ and are pure and counterless by [2, Proposition 7]. We can guess such a strategy σ^* and verify, using the numbers ν_r, that it is $\mathit{LimInf}(= -\infty)$-optimal for Max; similarly for π^* and Min. If $\mathcal{A}$ is an OC-MDP, all the above can be computed deterministically in time polynomial in $\|A\|$.

- (Step 2.) Fixing π^* in $\mathcal{A}$ we obtain a maximizing OC-MDP $\mathcal{A}^*$. Lemma 3.9 applied to $\mathcal{A}^*$ allows us to compute deterministically a bound $N \in \exp(\|\mathcal{A}'\|^{O(1)}) \cdot O(\log(1/\varepsilon))$ such that in $\mathcal{A}^*$, $\mathrm{Val}(\mathit{Term}, (r, j)) - \nu_r \leq \varepsilon$ for all $r \in Q$ and $j \geq N$. By Lemma 3.12 this N satisfies the requirements of step 2.

□

3.1. Bounding counter value N for maximizing OC-MDPs

Consider a maximizing OC-MDP $\mathcal{A} = (Q, \Delta, P)$. Recall the definition of ν_q from (1) and the notational convention introduced in Remark 3.1. Specifically, we have $\nu_q = \sup_\sigma \mathbb{P}_q^\sigma(\mathit{LimInf}(= -\infty))$ for all $q \in Q$. Given $\varepsilon > 0$, we show here how to obtain a computable (exponential) bound on a number N such that $\left|\mathrm{Val}(\mathit{Term}, (q, i)) - \nu_q\right| < \varepsilon$ for all $i \geq N$. We denote by T the set of all states q with $\nu_q = 1$, and we define the objective of reaching T as follows:

$$\mathit{Reach}_T := \{\omega \in \mathit{Run} \mid \mathrm{State}^{(i)}(\omega) \in T \text{ for some } i \geq 0\}.$$

Further, we define the objective $\neg\mathit{Reach}_T := \mathit{Run} \setminus \mathit{Reach}_T$.

Fact 3.2 (cf. [3, Proposition 3.2])**.** *The number ν_q is the maximal probability of reaching T from q (see Remark 3.1), i.e.,*

$$\nu_q = \mathrm{Val}(\mathit{Reach}_T, q) = \sup_\sigma \mathbb{P}_q^\sigma(\mathit{Reach}_T)\,.$$

Lemma 3.3. *For all $q \in Q$ and $i \geq 0$*

$$\nu_q \leq \mathrm{Val}(\mathit{Term}, (q, i)) \leq \sup_\sigma \mathbb{P}_{(q,i)}^\sigma(\mathit{Term} \cap \neg\mathit{Reach}_T) + \nu_q. \tag{2}$$

Proof. The first inequality is obvious. Because $\mathit{LimInf}(= -\infty) \cap \mathit{Run}((q, i)) \subseteq \mathit{Term} \cap \mathit{Run}((q, i))$, we have $\mathrm{Val}(\mathit{Term}, (q, i)) = 1$ for all $q \in T$, $i \geq 0$, from which the second inequality follows by an easy application of the union bound. Namely, for under strategy σ, the event of termination can be split into the event of terminating and not reaching T unioned with the event of terminating and reaching T. The probability of the latter event is clearly upper bounded by ν_q. □

To provide the promised bound on N we will prove an upper bound on $\sup_\sigma \mathbb{P}_{(q,i)}^\sigma(\mathit{Term} \cap \neg\mathit{Reach}_T)$ which decreases toward 0 exponentially fast in i. We will first define a suitably restricted class of OC-MDPs (we call them "rising"

OC-MDPs) and find such a bound using martingale theory (Lemma 3.8) for that restricted class. We then extend the results to all OC-MDPs (Lemma 3.9) by showing that for every OC-MDP, $\mathcal{A}$, there is a polynomially bigger "rising" OC-MDP, $\overline{\mathcal{A}}$, which "embedds" in it the states of the original OC-MDP, and preserves the rate with which $\sup_\sigma \mathbb{P}^\sigma_{(q,i)}(\mathit{Term} \cap \neg \mathit{Reach}_T)$ reaches 0 from those corresponding states. We will make this precise later.

To be able to use the martingale theory methods for rising OC-MDPs we need to guarantee that in each rising OC-MDP, under every pure counterless strategy, $\liminf_{i\to\infty} \mathrm{C}^{(i)}/i$ is almost surely positive. This value is sometimes called the mean-payoff, see also [3]. We now state the definition of a rising OC-MDP using two simple properties, and show that these two properties guarantee that the mean-payoff is almost surely positive.

Definition 3.4. A pure counterless strategy, σ, is *idling*, if there is a state $q \in Q$, such that $\mathbb{P}^\sigma_{(q,0)}\Big(\exists i > 0 : \mathrm{State}^{(i)} = q\Big) = 1$ and for all $i \geq 0$: $\mathbb{P}^\sigma_{(q,0)}\Big(\mathrm{State}^{(i)} = q \implies \mathrm{C}^{(i)} = 0\Big) = 1$.

A maximizing OC-MDP is called *rising* if $T = \{q \in Q \mid \nu_q = 1\} = \emptyset$ and no pure counterless strategy is idling.

Lemma 3.5. *Let $\mathcal{A} = (Q, \Delta, P)$ be a rising OC-MDP. Then for every pure counterless strategy, σ, and every $q \in Q$ we have $\mathbb{P}^\sigma_q\Big(\liminf_{i\to\infty} C^{(i)}/i > 0\Big) = 1$.*

Proof. Let us fix a pure counterless strategy σ. Because σ is counterless, there is a collection of disjoint subsets of Q, called ergodic sets, or bottom strongly connected components (BSCCs), in the standard theory of Markov chains, such that almost all runs end up visiting infinitely often exactly the states of some of the BSCCs. Let us focus, for a while, on a single BSCC $C \subseteq Q$. By standard results, for each pair of states $r, s \in C$ the play from r almost surely visits s, and the expected time to visit s from r is finite. As a consequence, there is a unique constant p such that $\mathbb{P}^\sigma_r\Big(\liminf_{i\to\infty} \mathrm{C}^{(i)}/i > 0\Big) = p$ for all $r \in C$, because $\liminf_{i\to\infty} \mathrm{C}^{(i)}/i > 0$ is a prefix-independent property. Moreover, we can use the result from [12, Theorem 3.2] which says that in a presence of $r \in C$ such that $\mathbb{P}^\sigma_r\Big(\liminf_{i\to\infty} \mathrm{C}^{(i)}/i > 0\Big) > 0$ there must also be some $s \in C$ such that $\mathbb{P}^\sigma_s\Big(\liminf_{i\to\infty} \mathrm{C}^{(i)}/i > 0\Big) = 1$. Thus either $p = 0$ or $p = 1$. Let us first prove by contradiction that $p = 1$, and then we shall consider the more general case where rather than assuming $r \in C$ for a BSCC C, we consider an arbitrary start state in the entire OC-MDP.

Assume that $p = 0$. By Fact 3.2, $T = \emptyset$ implies that $\mathbb{P}^{\sigma}_{r}\Big(\liminf_{i\to\infty} \mathrm{C}^{(i)} = -\infty\Big) = 0$ for all $r \in C$. It is easy to see that this implies that $\mathbb{P}^{\sigma}_{r}\Big(\liminf_{i\to\infty} \mathrm{C}^{(i)}/i \geq 0\Big) = 1$ for all $r \in C$ and all strategies σ. Due to our assumption of $p = 0$, $\mathbb{P}^{\sigma}_{r}\Big(\liminf_{i\to\infty} \mathrm{C}^{(i)}/i = 0\Big) = 1$ for all $r \in C$. Now Lemma 3.3 from [3] says: "For all q, the pure counterless strategies τ which satisfy $\mathbb{P}^{\tau}_{q}\Big(\liminf_{i\to\infty} \mathrm{C}^{(i)} = -\infty\Big) = 1$ are exactly those which satisfy $\mathbb{P}^{\tau}_{q}\Big(\liminf_{i\to\infty} \mathrm{C}^{(i)}/i \leq 0\Big) = 1$ and $\mathbb{P}^{\tau}_{(q,0)}\Big(\exists i : \mathrm{C}^{(i)} < 0\Big) > 0$." But we do not have any strategies of the first kind, so there is a state $r \in C$ such that $\mathbb{P}^{\sigma}_{(r,0)}\Big(\forall i : \mathrm{C}^{(i)} \geq 0\Big) = 1$. If $\mathbb{P}^{\sigma}_{(r,0)}\Big(\exists i > 0 : \mathrm{State}^{(i)} = r \wedge \mathrm{C}^{(i)} > 0\Big) > 0$ then because the expected time between two visits to r is finite, it can fairly easily be established that $\mathbb{P}^{\sigma}_{r}\Big(\liminf_{i\to\infty} \mathrm{C}^{(i)}/i > 0\Big) > 0$, which would contradict the assumption $p = 0$. Thus $\mathbb{P}^{\sigma}_{(r,0)}\Big(\forall i > 0 : \mathrm{State}^{(i)} = r \implies \mathrm{C}^{(i)} = 0\Big) = 1$ and σ is thus idling. But this is not possible because by assumption $\mathcal{A}$ is rising, so the assumption of $p = 0$ cannot be satisfied, and we have proved that $p = 1$.

Now consider an arbitrary state $q \in Q$. Because $\liminf_{i\to\infty} \mathrm{C}^{(i)}/i > 0$ is prefix-independent, and almost every run from q reaches some BSCC, where $\liminf_{i\to\infty} \mathrm{C}^{(i)}/i > 0$ is satisfied almost surely, we have $\mathbb{P}^{\sigma}_{q}\Big(\liminf_{i\to\infty} \mathrm{C}^{(i)}/i > 0\Big) = 1$. □

Now we define a suitable submartingale for a given rising OC-MDP, and use Azuma's inequality to show that $\sup_{\sigma} \mathbb{P}^{\sigma}_{(q,i)}(\mathit{Term} \cap \neg \mathit{Reach}_{T})$ decreases to 0 exponentially fast in i. Recall that a stochastic process $m^{(0)}, m^{(1)}, \ldots$ is a submartingale if, for all $i \geq 0$, $\mathbb{E}\Big[|m^{(i)}|\Big] < \infty$, and $\mathbb{E}\Big[m^{(i+1)} \mid m^{(1)}, \ldots, m^{(i)}\Big] \geq m^{(i)}$ almost surely. If we further assume that $|m^{(i+1)} - m^{(i)}| \leq c$ almost surely for all $i \geq 0$, we can apply the *Azuma-Hoeffding inequality*[5], which says that the following holds for all $t > 0$ and $n \geq 0$:

$$\mathbb{P}\Big(m^{(n)} - m^{(0)} \leq t\Big) \quad \leq \quad \exp\left(\frac{-t^2}{2nc^2}\right) \tag{3}$$

Let $\mathcal{A} = (Q, \Delta, P)$ be a rising OC-MDP. Since $\mathcal{A}$ is rising, the mean-payoff (i.e., the average change of the counter per transition) is almost surely positive for all pure counterless strategies. Since there are only finitely many pure counterless

[5]In the literature (see, e.g. [13]), the Azuma-Hoeffding inequality is usually stated for martingales and supermartingales where is takes the form $\mathbb{P}\Big(m^{(n)} - m^{(0)} \geq t\Big) \leq \exp(-t^2/2nc^2)$. Inequality (3) is obtained just by realizing that if $m^{(0)}, m^{(1)}, \ldots$ is a submartingale, then $-m^{(0)}, -m^{(1)}, \ldots$ is a supermartingale.

strategies, there is even a fixed bound $x > 0$ such that the mean payoff is larger than x almost surely. This means that after performing i transitions, the counter should increase at least by $i \cdot x$ on average. Hence, one might be tempted to define

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

and try to prove that $m^{(0)}, m^{(1)}, \ldots$ is a submartingale. Unfortunately, this does not work, because some control states may not allow to increase the counter by x or more. A similar problem was encountered previously in [6] in the context of purely probabilistic OC automata, and the difficulty was overcome by employing "artificial" additive constants that compensate the difference among the individual control states. We show that a similar trick works also in our setting. That is, we aim at designing a submartingale of the following form:

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

Here z_q is a suitable constant that depends only on q. However, it is not clear whether the constants z_q can be chosen so that $m^{(0)}, m^{(1)}, \ldots$ becomes a submartingale, and what is the size of these constants if they exist. This problem is solved simply by observing that the defining property of a submartingale (see above) immediately gives a system of linear inequality constraints that should be satisfied by z_q. For example, suppose that $\mathrm{C}^{(i)} = j$ and $\mathrm{State}^{(i)} = q$ where $q \in Q_P$. For *every* Max strategy, we would like to have that $\mathbb{E}\left[m^{(i+1)} \mid m^{(i)}\right] \geq m^{(i)}$. This means to ensure that this inequality is satisfied for every outgoing transition of (q, j), i.e., for every $(q, k, r) \in \Delta$ we wish to have

$$\mathbb{E}\left[m^{(i+1)} \mid m^{(i)}\right] \quad = \quad (j + k) + z_r - (i + 1) \cdot \overline{x} \quad \geq \quad m^{(i)} \quad = \quad j + z_q - i \cdot \overline{x}.$$

This yields $z_q \leq -x + k + z_r$. Note that if q is stochastic, we need to consider the "weighted sum" of the outgoing transition of (q, j) instead. Thus, we obtain the system of linear inequalities of Figure 1.

Now we show that the linear system of inequalities given in Figure 1 has a non-negative rational solution, and derive a bound on its size. Then, we take this solution, define the associated submartingale, and use Azuma's inequality to derive the desired result.

Lemma 3.6. *Let $\mathcal{A} = (Q, \Delta, P)$ be a rising OC-MDP. Then there is a non-negative rational solution $(\overline{x}, (\overline{z}_q)_{q \in Q}) \in \mathbb{Q}^{|Q|+1}$ to $\mathcal{L}$, such that $\overline{x} > 0$. (The binary encoding size of the solution is polynomial in $\|\mathcal{A}\|$.)*

$$
\begin{aligned}
& z_q \le -x + k + z_r && \text{for all } q \in Q_\top \text{ and } (q,k,r) \in \Delta,\\
& z_q \le -x + \textstyle\sum_{(q,k,r)\in\Delta} P((q,k,r)) \cdot (k + z_r) && \text{for all } q \in Q_P,\\
& x > 0.
\end{aligned}
$$

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

Proof. We first prove that there is some non-negative solution to $\mathcal{L}$ with $\overline{x} > 0$. The bound on size then follows by standard facts about linear programming. To find a solution, we will use optimal values for minimizing *discounted total reward* in $\mathcal{A}$. For every discount factor, λ, $0 < \lambda < 1$, and a strategy, τ, we denote the discounted total reward, starting under τ, by $DTR_q^\lambda(\tau) \coloneqq \sum_{i\ge 0} \lambda^i \cdot \mathbb{E}_q^\tau\left[\mathrm{C}^{(i+1)} - \mathrm{C}^{(i)}\right]$, and set $DTR_q^\lambda(*) \coloneqq \inf_\tau DTR_q^\lambda(\tau)$. We prove that there is some λ, such that setting $\overline{z_q} \coloneqq DTR_q^\lambda(*)$ and

$$
\begin{aligned}
\overline{x} \coloneqq \min\big(&\{k + DTR_r^\lambda(*) - DTR_q^\lambda(*) \mid q \in Q_\top, (q,k,r) \in \Delta\}\\
&\cup \{P((q,k,r)) \cdot \left(k + DTR_r^\lambda(*) - DTR_q^\lambda(*)\right) \mid q \in Q_P, (q,k,r) \in \Delta\}\big)
\end{aligned}
$$

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

Now we proceed in more detail. First we choose the right λ. Lemma 3.5 and our assumptions guarantee that $\mathbb{P}_q^\tau\left(\liminf_{i\to\infty} \mathrm{C}^{(i)}/i > 0\right) = 1$ for every pure counterless strategy τ. Thus $\sum_{i\ge 0} \cdot \mathbb{E}_q^\tau\left[\mathrm{C}^{(i+1)} - \mathrm{C}^{(i)}\right] = \infty$, and hence for every such τ there is a $\Lambda_\tau < 1$ such that $DTR_q^\lambda(\tau) > 0$ for all $q \in Q$ and $\lambda \ge \Lambda_\tau$. There are only finitely many pure counterless strategies, and we choose our λ to be $\lambda \coloneqq \max_\tau \Lambda_\tau < 1$.

Having fixed the λ above, we now prove that there is a pure counterless strategy σ, such that $DTR_q^\lambda(*) = DTR_q^\lambda(\sigma)$ for all q. By standard results (e.g., [16]), translated from the terminology of MDPs with rewards to that of OC-MDPs, for a fixed state, q, there is always a pure counterless strategy σ_q such that $DTR_q^\lambda(\sigma_q) = DTR_q^\lambda(*)$. Moreover, this strategy has to play optimally in successors of q as well, thus there is in fact a single pure counterless strategy σ such that for all q: $DTR_q^\lambda(\sigma) = DTR_q^\lambda(*)$.

Finally, $\overline{x} > 0$, because for all $q \in Q_P$

$$\begin{aligned} DTR_q^\lambda(\sigma) &= \sum_{i\geq 0} \lambda^i \cdot \mathbb{E}_q^\sigma\left[\mathrm{C}^{(i+1)} - \mathrm{C}^{(i)}\right] \\ &= \sum_{(q,k,r)\in\Delta} P((q,k,r)) \cdot \left(k + \lambda \cdot \sum_{i\geq 0} \lambda^i \cdot \mathbb{E}_r^\sigma\left[\mathrm{C}^{(i+1)} - \mathrm{C}^{(i)}\right]\right) \\ &= \sum_{(q,k,r)\in\Delta} P((q,k,r)) \cdot \left(k + \lambda \cdot DTR_r^\lambda(\sigma)\right) \\ &< \sum_{(q,k,r)\in\Delta} P((q,k,r)) \cdot \left(k + DTR_r^\lambda(\sigma)\right), \end{aligned}$$

the last inequality following from $DTR_r^\lambda(\sigma) > 0$ for all $r \in Q$; and similarly for all $q \in Q_\top$ and $(q,k,r) \in \Delta$

$$\begin{aligned} DTR_q^\lambda(\sigma) &= \sum_{i\geq 0} \lambda^i \cdot \mathbb{E}_q^\sigma\left[\mathrm{C}^{(i+1)} - \mathrm{C}^{(i)}\right] \\ &\leq k + \lambda \cdot \sum_{i\geq 0} \lambda^i \cdot \mathbb{E}_r^\sigma\left[\mathrm{C}^{(i+1)} - \mathrm{C}^{(i)}\right] = k + \lambda \cdot DTR_r^\lambda(\sigma) < k + DTR_r^\lambda(\sigma). \end{aligned}$$

Here the first inequality follows from the fact that σ minimizes the discounted total reward. □

Given the solution $(\overline{x}, (\overline{z}_q)_{q\in Q}) \in \mathbb{Q}^{|Q|+1}$ from Lemma 3.6, we define a sequence of random variables $\{m^{(i)}\}_{i\geq 0}$ by setting

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

We shall now show that $m^{(i)}$ defines a submartingale.

Lemma 3.7. *Let $\mathcal{A} = (Q, \Delta, P)$ be a rising OC-MDP and $\{m^{(i)}\}_{i\geq 0}$ defined as above. Under an arbitrary strategy τ 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 $\mathrm{C}^{(j)}(\omega)$ are the same for all $\omega \in Run(u)$. We denote these common values by $\mathrm{C}^{(j)}(u)$, and similarly for $\mathrm{State}^{(j)}(u)$ and $m^{(j)}(u)$. If $\mathrm{C}^{(j)}(u) = 0$ for some $j \leq i$, then $m^{(i+1)}(\omega) = m^{(i)}(\omega)$ for every $\omega \in Run(u)$. Thus $\mathbb{E}_{(q,n)}^\tau\left[m^{(i+1)} \mid Run(u)\right] =$

$m^{(i)}(u)$. Otherwise, consider the last configuration, (r, l), of u. For every possible successor, (r', l'), set

$$p_{(r',l')} \coloneqq \begin{cases} \tau(u)((r,l) \to (r',l')) & \text{if } r \in Q_\top, \\ \mathit{Prob}((r,l) \to (r',l')) & \text{if } r \in Q_P. \end{cases}$$

Then

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

This allows us to derive the following:

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

□

Now we have prepared all that we need to bound $\sup_\sigma \mathbb{P}^{\sigma}_{(q,i)}(\mathit{Term})$ for rising OC-MDPs.

Lemma 3.8. *Given a rising 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 $q \in Q$*

$$\sup_\sigma \mathbb{P}^{\sigma}_{(q,i)}(\mathit{Term}) \leq \frac{c^i}{1-c}.$$

Moreover, $c \in \exp(1/2^{\|\mathcal{A}\|^{O(1)}})$ and $h \in \exp(\|\mathcal{A}\|^{O(1)})$.

Proof. Denote by Term_j the event of terminating after *exactly* j steps. Further set $\overline{z}_{\max} \coloneqq \max_{q\in Q}\overline{z}_q - \min_{q\in Q}\overline{z}_q$, and assume that $\mathrm{C}^{(0)} \geq \overline{z}_{\max}$. Then the event Term_j implies that $m^{(j)} - m^{(0)} = \overline{z}_{\mathrm{State}^{(j)}} - j\cdot\overline{x} - \mathrm{C}^{(0)} - \overline{z}_{\mathrm{State}^{(0)}} \leq -j\cdot\overline{x}$. Finally, observe that we can bound the one-step change of the submartingale value by $\overline{z}_{\max} + \overline{x} + 1$. Using the Azuma-Hoeffding inequality for the submartingale $\{m^{(n)}\}_{n\geq 0}$ (see, e.g., Theorem 12.2.3 in [13]), we thus obtain the following bound for every strategy σ and initial configuration (q, i) with $i \geq \overline{z}_{\max}$:

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

We choose $c := \exp\left(\frac{-\vec{x}^2}{2\cdot(\bar{z}_{\max}+\bar{x}+1)}\right) < 1$ and $h := \lceil \bar{z}_{\max} \rceil$, and observe that for all $q \in Q$, $i \geq h$:

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

The given bounds on c and h are easy to check, and the detailed computation can be found in Section A.3. □

As a final step, we extend the results to the general case of (not necessarily rising) OC-MDPs.

Lemma 3.9. *Given a maximizing OC-MDP, $\mathcal{A}' = (Q', \Delta', P')$, one can compute a rational constant $c < 1$, and an integer $h \geq 0$ such that for all $i \geq h$ and $q \in Q$*

$$\sup_{\sigma} \mathbb{P}^{\sigma}_{(q,i)}(\mathit{Term} \cap \neg \mathit{Reach}_T) \leq \frac{c^i}{1-c}.$$

Moreover, $c \in \exp(1/2^{\|\mathcal{A}'\|^{O(1)}})$ and $h \in \exp(\|\mathcal{A}'\|^{O(1)})$. As a consequence, a number N such that $\left|\mathrm{Val}(\mathit{Term}, (q,i)) - \sup_{\sigma} \mathbb{P}^{\sigma}_{q}(\mathit{LimInf}(= -\infty))\right| < \varepsilon$ for all $q \in Q'$ and $i \geq N$ satisfies $N \leq \max\{h, \lceil \log_c(\varepsilon \cdot (1-c)) \rceil\} \in \exp(\|\mathcal{A}'\|^{O(1)}) \cdot O(\log(1/\varepsilon))$.

Proof. The heart of the proof is a reduction which computes a polynomially bigger rising OC-MDP $\bar{\mathcal{A}} = (\bar{Q}, \bar{\Delta}, \bar{P})$ from $\mathcal{A}'$, uses the algorithm from Lemma 3.8 to compute the bounds c and h for $\bar{\mathcal{A}}$, and returns the very same numbers for $\mathcal{A}'$. The reduction itself is in two steps, first computing an OC-MDP $\mathcal{A} = (Q, \Delta, P)$ from $\mathcal{A}'$ such that $\mathrm{Val}(\mathit{LimInf}(= -\infty), q) < 1$ for all $q \in Q$ in $\mathcal{A}$, and then $\bar{\mathcal{A}}$ from $\mathcal{A}$.

The first step, from $\mathcal{A}'$ to $\mathcal{A}$ is easier. Recall that we called $T = T(Q)$ the set of all $q \in Q$ such that $\mathrm{Val}(\mathit{LimInf}(= -\infty), q) = 1$. Here we use it also to denote the analogous subset $T(Q')$ of Q' of all states $q \in Q'$ such that $\mathrm{Val}(\mathit{LimInf}(= -\infty), q) = 1$ in $\mathcal{A}'$. Theorem 3.1 from [3] guarantees that we can compute the set $T(Q')$ in time polynomial in $\|\mathcal{A}'\|$. Then we set $Q := (Q' \setminus T(Q')) \cup \{\mathsf{trap}\}$, with $Q_P = Q'_P \setminus T(Q')$. The state trap has a unique outgoing rule in Δ: $(\mathsf{trap}, +1, \mathsf{trap})$. The rest of the rules in Δ are derived from Δ' by redirecting all rules ending in $T(Q')$ to trap. P' is derived from $\bar{P}$ accordingly. It is easy to see that $T(Q) = \emptyset$, because $T(Q') \cap Q = \emptyset$ and by construction, $T(Q) \subseteq T(Q')$.

A technique to achieve the second step was already partially developed in our previous work [3], where we used the term "decreasing" for rising strategies.

There we gave a construction which preserves the property of optimal termination probability being = 1. We in fact can establish that a similar construction preserves the exact termination value. Because the idea is not new, we leave details to Section A.2. The important properties of $\bar{\mathcal{A}}$ are stated in the following lemma, the proof of which can be found in Section A.2, along with the formal definition of $\bar{\mathcal{A}}$.

Lemma 3.10. *There is a deterministic polynomial-time algorithm which given a maximizing OC-MDP, $\mathcal{A} = (Q, \Delta, P)$, computes another maximizing OC-MDP, $\bar{\mathcal{A}} = (\bar{Q}, \bar{\Delta}, \bar{P})$, and a map $f : Q \to \bar{Q}$ satisfying:*

- $\|\bar{\mathcal{A}}\| \in O(\|\mathcal{A}\|^4)$.
- *There are no idling pure counterless strategies in $\bar{\mathcal{A}}$.*
- $\mathrm{Val}(\mathit{Term}, (q, i)) = \mathrm{Val}(\mathit{Term}, (f(q), i))$ *for all $q \in Q$ and $i \geq 0$.*
- *If* $\mathrm{Val}(\mathit{LimInf}(= -\infty), q) < 1$ *for all $q \in Q$ in $\mathcal{A}$, then $\bar{\mathcal{A}}$ is rising.*

In particular, note that $\bar{\mathcal{A}}$ obtained from our $\mathcal{A}$ is rising. Now let $q \in Q'$ be a state of $\mathcal{A}'$, such that $q \notin T(Q')$. We know that $\sup_\sigma \mathbb{P}^\sigma_{(q,i)}(\mathit{Term} \cap \neg \mathit{Reach}_T)$ in $\mathcal{A}'$ equals $\sup_\sigma \mathbb{P}^\sigma_{(q,i)}(\mathit{Term})$ in $\mathcal{A}$, which in turn equals $\sup_\sigma \mathbb{P}^\sigma_{(f(q),i)}(\mathit{Term})$ in $\bar{\mathcal{A}}$. Note that $\|\bar{\mathcal{A}}\| \in \|\mathcal{A}'^{O(1)}\|$. Applying Lemma 3.8 to $\bar{\mathcal{A}}$ finishes the proof of the first part of Lemma 3.9.

In the second part the inequality $N \leq \max\{h, \lceil \log_c(\varepsilon \cdot (1-c)) \rceil\}$ is an easy computation. Verifying that $\lceil \log_c(\varepsilon \cdot (1-c)) \rceil \in \exp(\|\mathcal{A}'\|^{O(1)})$ is also easy and can be found in Section A.3. □

Also, as an immediate consequence of Lemma 3.3 and Lemma 3.9 we obtain the following:

Corollary 3.11. *For every $q \in Q$, $v_q = \lim_{i\to\infty} \mathrm{Val}(\mathit{Term}, (q, i))$.*

3.2. Bounding N for general SSGs

By [2, Proposition 7], player Min always has an optimal pure counterless strategy, π^*, such that

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

By fixing the choices of π^* in $\mathcal{A}$ we obtain a maximizing OC-MDP, $\mathcal{A}^* = (Q^*, \delta^*, P^*)$, where $Q^*_P = Q_P \cup Q_\perp$, $Q^*_\top = Q_\top$, $\delta^* := \{(q, k, r) \in \delta \mid q \in Q_P \cup Q_\top \vee \pi^*(q) = r\}$, and P^* is the unique (for $\mathcal{A}^*$) extension of P to states from $Q_\perp$.

Lemma 3.12. *Let $\mathcal{A} = (Q, \Delta, P)$ be an OC-SSG, π^* a LimInf$(= -\infty)$-optimal strategy for Min, and $\mathcal{A}^*$ the minimizing OC-MDP given by fixing π^* in $\mathcal{A}$ as described above. Then for all $q \in Q$:*

$$\forall i \geq 0 : \lim_{j\to\infty} \mathrm{Val}_{\mathcal{A}}(\mathit{Term}, (q, j)) \leq \mathrm{Val}_{\mathcal{A}}(\mathit{Term}, (q, i)) \leq \mathrm{Val}_{\mathcal{A}^*}(\mathit{Term}, (q, i)). \quad (4)$$

$$\lim_{j\to\infty} \mathrm{Val}_{\mathcal{A}}(\mathit{Term}, (q, j)) = \lim_{j\to\infty} \mathrm{Val}_{\mathcal{A}^*}(\mathit{Term}, (q, j)). \quad (5)$$

Proof. Since for all j we have $\mathrm{Val}_{\mathcal{A}}(\mathit{Term}, (q, j+1)) \leq \mathrm{Val}_{\mathcal{A}}(\mathit{Term}, (q, j))$,, we obtain the first inequality in (4). The second inequality in (4) follows from the fact that in $\mathcal{A}^*$ we restricted the possible moves of Min. The "$\leq$" direction in (5) follows directly from (4), and the other direction is obtained as follows:

$$\begin{array}{lcll} \lim_{i\to\infty} \mathrm{Val}_{\mathcal{A}}(\mathit{Term}, (q, i)) & \geq & \mathrm{Val}_{\mathcal{A}}(\mathit{LimInf}(= -\infty), q) & \text{(immediate)} \\ & = & \mathrm{Val}_{\mathcal{A}^*}(\mathit{LimInf}(= -\infty), q) & \text{(immediate)} \\ & = & \lim_{i\to\infty} \mathrm{Val}_{\mathcal{A}^*}(\mathit{Term}, (q, i)) & \text{(by Corollary 3.11)} \end{array}$$

□

Corollary 3.13. *For every control state q of an OC-SSG $\mathcal{A}$ we have that*

$$\lim_{i\to\infty} \mathrm{Val}_{\mathcal{A}}(\mathit{Term}, (q, i)) = \mathrm{Val}_{\mathcal{A}}(\mathit{LimInf}(= -\infty), q).$$

3.3. Analyzing a Finite Segment of Configurations

Lemma 3.14. *There is a nondeterministic algorithm*[6] *that given an OC-SSG, $\mathcal{A} = (Q, \Delta, P)$, and a rational $\varepsilon > 0$ as input, and, in addition, given the following precomputed values:*

- *$\nu_q = \mathrm{Val}(\mathit{LimInf}(= -\infty), q)$ for every $q \in Q$*
- *an integer $N \geq 0$ such that $0 \leq \mathrm{Val}(\mathit{Term}, (q, i)) - \nu_q \leq \varepsilon$ for all $q \in Q$ and $i \geq N$,*
- *and a pair of strategies (σ^*, π^*) for Max and Min which are optimal for LimInf$(= -\infty)$ in all $q \in Q$;*

computes the following output:

[6] Again, see footnote 4 for a precise explanation of what we mean by a nondeterministic algorithm in this context.

- *a number* $\nu_{(q,i)}$ *for each* $q \in Q$ *and* $i \leq N$ *such that* $0 \leq \mathrm{Val}(\mathit{Term}, (q,i)) - \nu_{(q,i)} \leq \varepsilon$,
- *and a pair of strategies* $(\overline{\sigma}, \overline{\pi})$ *for Max and Min, respectively, which are* ε*-optimal for termination in all configurations.*

The algorithm runs in time polynomial in $N \cdot \|\mathcal{A}\|$*. Furthermore, if* $\mathcal{A}$ *is an OC-MDP then the algorithm is deterministic.*

Proof. The first idea is to analyze the following SSG, $\mathcal{G}$, which is essentially $\mathcal{A}$ restricted to configurations with counter value between 0 and N. The set of states of $\mathcal{G}$ is $\{(q,i) \mid q \in Q, 0 \leq i \leq N\} \cup \{s_0, s_1\}$. The ownership of the states of the form (q,i), $0 < i < N$ is the same as in $\mathcal{A}$, the states s_0, s_1 and (q,i) for $q \in Q$, $i \in \{0, N\}$ are stochastic. For $0 < i < N$, there is a transition $(q,i) \rightsquigarrow (r,j)$ iff $(q, j-i, r) \in \delta$. Probabilities of these transitions, where applicable, are derived from P. Vertices of the form $(q,0)$, and the state s_0 have only one transition, to s_0. Vertices of the form (q,N) have transitions to both s_0 and s_1, and the state s_1 has only the self-loop transition. The probability of a transition $(q,N) \rightsquigarrow s_0$ equals $\mathrm{Val}(\mathit{Term}, (q,N))$ for all q.

Clearly we have $\sup_\sigma \inf_\pi \mathbb{P}^{\sigma,\pi}_{(q,i)}(\text{reach } s_0) = \mathrm{Val}(\mathit{Term}, (q,i))$ for all $q \in Q$ and $i \leq N$. The problem is that the transition probabilities from (q,N) in $\mathcal{G}$ are unknown (and may even be irrational). We will not actually construct $\mathcal{G}$. To use such reachability analysis for approximating the termination values we have to switch to a slightly perturbed SSG, $\mathcal{G}'$.

$\mathcal{G}'$ is almost identical to $\mathcal{G}$: it has the same sets of states and transitions. The only difference is that in $\mathcal{G}'$ the probability of a transition $(q,N) \rightsquigarrow s_0$ equals ν_q for every q (and the probability of $(q,N) \rightsquigarrow s_1$ changes appropriately to make the sum 1). Observe that since $\nu_q \leq \mathrm{Val}(\mathit{Term}, (q,N))$, for every (q,i) where $i \leq N$:

$$\sup_\sigma \inf_\pi \mathbb{P}^{\sigma,\pi}_{(q,i)}(\text{reach } s_0 \text{ in } \mathcal{G}') \leq \sup_\sigma \inf_\pi \mathbb{P}^{\sigma,\pi}_{(q,i)}(\text{reach } s_0 \text{ in } \mathcal{G}) = \mathrm{Val}(\mathit{Term}, (q,i)).$$

On the other hand, by our assumption on the values ν_q and N,

$$\sup_\sigma \inf_\pi \mathbb{P}^{\sigma,\pi}_{(q,i)}(\text{reach } s_0 \text{ in } \mathcal{G}) - \varepsilon \leq \sup_\sigma \inf_\pi \mathbb{P}^{\sigma,\pi}_{(q,i)}(\text{reach } s_0 \text{ in } \mathcal{G}')\,.$$

Thus $\mathrm{Val}(\mathit{Term}, (q,i)) - \sup_\sigma \inf_\pi \mathbb{P}^{\sigma,\pi}_{(q,i)}(\text{reach } s_0 \text{ in } \mathcal{G}') \leq \varepsilon$, and we may output $\nu_{(q,i)} := \sup_\sigma \inf_\pi \mathbb{P}^{\sigma,\pi}_{(q,i)}(\text{reach } s_0 \text{ in } \mathcal{G}')$. By standard results, see, e.g., [7], such reachability values have a binary encoding polynomial in $\|\mathcal{G}'\|$, and after a memoryless optimal strategy (having size polynomial in $\|\mathcal{G}'\|$) is guessed, the values

can be computed in time polynomial in $\|\mathcal{G}'\|$. If $\mathcal{A}$ is an OC-MDP, then $\mathcal{G}'$ is an MDP, and for MDPs the reachability values, and optimal strategies, can be computed in deterministic polynomial time. Let us suppose we have computed the optimal strategies σ_R, π_R for reachability in $\mathcal{G}'$. The resulting strategy $\bar{\sigma}$ for the given OC-SSG $\mathcal{A}$ is defined as follows: In configurations with counter value between 0 and N it plays according to the optimal reachability strategy of Max in $\mathcal{G}'$. Once a configuration with a counter value above N is visited it switches to playing as σ^* forever, where σ^* is the optimal stategy we assume we are given for $\mathit{LimInf}(= -\infty)$. Now for all configurations (q, i), $0 \leq i < N$, and strategies π for Min, the number $\nu_{(q,i)}$ gives a lower bound on the probability that under $(\bar{\sigma}, \pi)$ a run either terminates without exceeding counter value N, or hits some (r, N) and then satisfies $\mathit{LimInf}(= -\infty)$. This probability itself is a lower bound for the probability that a run either terminates without exceeding counter value N, or hits some (r, N) and then terminates, which is in other words the probability of termination. This means that $\bar{\sigma}$ is ε-optimal, because $\mathrm{Val}(\mathit{Term}, (q, i)) - \nu_{(q,i)} \leq \varepsilon$.

Analogously we define the strategy $\bar{\pi}$. Consider again some (q, i), $0 \leq i < N$, and σ for Max. The number $\nu_{(q,i)}$ gives now an upper bound on the probability that under $(\sigma, \bar{\pi})$ a run either terminates without exceeding counter value N, or hits some (r, N) and then satisfies $\mathit{LimInf}(= -\infty)$. From the properties of N, this probability is by at most ε lower than the probability of termination. Because $\nu_{(q,i)} \leq \mathrm{Val}(\mathit{Term}, (q, i))$ we obtain that also $\bar{\pi}$ is ε-optimal. □

4. Conclusions

We have shown that one can ϵ-approximate the termination value for OC-MDP (and for OC-SSG) termination games, and compute ϵ-optimal strategies for them, in exponential time (and in nondeterministic exponential time, respectively).

Our results leave open several intriguing problems. An obvious remaining open problem is to obtain better complexity bounds. In particular, we know of no non-trivial lower bounds for OC-MDP approximation problems, and it remains possible that approximation of the value for OC-MDPs can be computed in polynomial time. 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 the computability of 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 [2, 3].

References

[1] N. Berger, N. Kapur, L. J. Schulman, and V. Vazirani. Solvency Games. In *Proc. of FSTTCS'08*, 2008.

[2] T. Brázdil, V. Brožek, and K. Etessami. One-Counter Simple Stochastic Games. In *Proc. of FSTTCS'10*, pages 108–119, 2010.

[3] T. Brázdil, V. Brožek, K. Etessami, A. Kučera, and D. Wojtczak. One-Counter Markov Decision Processes. In *ACM-SIAM SODA*, pages 863–874, 2010. Full tech report: CoRR, abs/0904.2511, 2009. http://arxiv.org/abs/0904.2511.

[4] T. Brázdil, V. Brožek, V. Forejt, and A. Kučera. Reachability in recursive Markov decision processes. In *Proc. 17th Int. CONCUR*, pages 358–374, 2006.

[5] T. Brázdil, V. Brožek, A. Kučera, and J. Obdržálek. Qualitative Reachability in stochastic BPA games. In *Proc. 26th STACS*, pages 207–218, 2009.

[6] T. Brázdil, S. Kiefer, and A. Kučera. Efficient analysis of probabilistic programs with an unbounded counter. *CoRR*, abs/1102.2529, 2011.

[7] A. Condon. The Complexity of Stochastic Games. *Inform. and Comput.*, 96:203–224, 1992.

[8] C. Courcoubetis and M. Yannakakis. Markov decision processes and regular events. *IEEE Trans. Automat. Control*, 43(10):1399–1418, 1998.

[9] K. Etessami, D. Wojtczak, and M. Yannakakis. 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)*, pages 243–253, 2008.

[10] K. Etessami and M. Yannakakis. Recursive Markov decision processes and recursive stochastic games. In *Proc. 32nd ICALP*, pages 891–903, 2005.

[11] K. Etessami and M. Yannakakis. Efficient qualitative analysis of classes of recursive Markov decision processes and simple stochastic games. In *Proc. of 23rd STACS'06*. Springer, 2006.

[12] H. Gimbert and F. Horn. Solving Simple Stochastic Tail Games. In *ACM-SIAM Symposium on Discrete Algorithms (SODA10)*, pages 847–862, 2010.

[13] G. R. Grimmett and D. R. Stirzaker. *Probability and Random Processes*. Oxford U. Press, 2nd edition, 1992.

[14] J. Lambert, B. Van Houdt, and C. Blondia. A policy iteration algorithm for markov decision processes skip-free in one direction. In *ValueTools*, Brussels, Belgium, 2007. ICST.

[15] D. A. Martin. The Determinacy of Blackwell Games. *The Journal of Symbolic Logic*, 63(4):1565–1581, December 1998.

[16] M. L. Puterman. *Markov Decision Processes*. J. Wiley and Sons, 1994.

[17] L. B. White. A new policy iteration algorithm for Markov decision processes with quasi birth-death structure. *Stochastic Models*, 21:785–797, 2005.

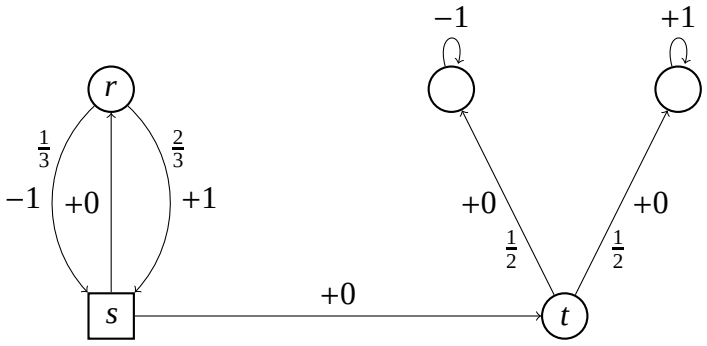

Figure 2: An OC-MDP where Player Max does not have optimal strategies for termination. Signed numbers represent counter increments, unsigned are probabilities of transitions.

A. Appendix

A.1. Non-existence of Optimal Strategies for Termination

In the following example we show that even in the special case of OC-MDPs there may not be any optimal strategies for maximizing the termination values. More precisely, there is a maximizing OC-MDP, $\mathcal{A}$, and (infinitely many) configurations (s, i) such that for all strategies σ: $\mathbb{P}^{\sigma}_{(s,i)}(\mathit{Term}) < \mathrm{Val}(\mathit{Term}, (q, i))$.

Example A.1. Consider the maximizing OC-MDP, $\mathcal{A}$, given in Figure 2. In the graph, round nodes represent stochastic states, the only square node is a state of Player Max, s. The arrows represent the rules, with signed numbers representing the increments, and non-signed the probabilities. For example the arrow from s to r represents the rule $(s, 0, r)$, whereas the right arrow from r to s represents the rule $(r, +1, s)$, which has probability $P(r, 1, s) = 2/3$.

Claim 1. *If the rule* $(s, 0, t)$ *is removed then* $\mathrm{Val}(\mathit{Term}, (s, i)) = 2^{-i}$.

Proof. Observe that there is only one strategy when the rule above is removed. We will omit writing its name. Clearly $\mathbb{P}_{(s,0)}(\mathit{Term}) = 1 = 2^0$. Further, the assignment $x := \mathbb{P}_{(s,1)}(\mathit{Term})$ is the least non-negative solution of the equation $x = \frac{1}{3} + \frac{2x^2}{3}$, which is $\frac{1}{2}$. Finally, $\mathbb{P}_{(s,i)}(\mathit{Term}) = \frac{1}{3} \cdot \mathbb{P}_{(s,i-1)}(\mathit{Term}) + \frac{2}{3} \cdot \mathbb{P}_{(s,i+1)}(\mathit{Term})$. Given the initial conditions for $i = 0, 1$, we obtain $\mathbb{P}_{(s,i)}(\mathit{Term}) = 2^{-i}$ as a unique solution of this recurrence. □

Claim 2. $\mathrm{Val}(\mathit{Term}, (s, 1)) = \frac{3}{4}$.

Proof. First we prove that $\mathrm{Val}(\mathit{Term}, (s,1)) \geq \frac{3}{4}$. For any n consider the pure strategy, σ_n, given for all histories ending in (s,i) by $\sigma_n(u)((s,i) \to (r,i)) = 1$ if $i < n$ and $\sigma_n(u)((s,i) \to (t,i)) = 1$ if $i \geq n$. Set

$$p_i \coloneqq \mathbb{P}^{\sigma_i}_{(s,1)}(\text{reach } (s,i)) .$$

Observe that p_i stays the same number if we define it using any σ_n with $n \geq i$, and that $1 - p_i = \mathbb{P}^{\sigma_i}_{(s,1)}(\text{terminate before reaching } (s,i))$. Moreover, $p_1 = 1$ and $p_{i+1} \coloneqq \frac{2}{3} \cdot (p_i + (1 - p_i) \cdot p_{i+1})$. This uniquely determines that $p_i = \frac{2^{i-1}}{2^i - 1}$. Note that $\lim_{i\to\infty} p_i = \frac{1}{2}$. Finally, observe that

$$\mathbb{P}^{\sigma_n}_{(s,1)}(\mathit{Term}) = (1 - p_n) + p_n \cdot \frac{1}{2} .$$

Thus, as $n \to \infty$ the probability of termination under σ_n approaches $\frac{3}{4}$.

Now we prove that $\mathrm{Val}(\mathit{Term}, (s,1)) \leq \frac{3}{4}$ by proving that $\mathbb{P}^{\sigma}_{(s,1)}(\text{terminate}) \leq \frac{3}{4}$ for every σ. Consider the following probabilities:

$$\begin{aligned} p_a &\coloneqq \mathbb{P}^{\sigma}_{(s,1)}(\text{terminate without visiting } t) , \\ p_b &\coloneqq \mathbb{P}^{\sigma}_{(s,1)}(\text{terminate after visiting } t) , \\ p_c &\coloneqq \mathbb{P}^{\sigma}_{(s,1)}(\text{visit } t) . \end{aligned}$$

Clearly $p_b = \frac{p_c}{2}$. Due to the first Claim, applied to $i = 1$, we also have that $p_a \leq \frac{1}{2}$. Finally, $p_a + p_c \leq 1$ since the events are disjoint. We conclude that

$$\mathbb{P}^{\sigma}_{(s,1)}(\mathit{Term}) = p_a + p_b \leq p_a + \frac{1}{2} \cdot (1 - p_a) = \frac{1}{2} \cdot p_a + \frac{1}{2} \leq \frac{3}{4}.$$

□

Claim 3. *For all* $i \geq 0$, $\mathrm{Val}(\mathit{Term}, (s,i)) = \frac{2^i+1}{2^{i+1}}$.

Proof. The case $i = 0$ is trivial, and $i = 1$ is by the previous Claim. Observe that $\mathrm{Val}(\mathit{Term}, (s,i)) \geq \frac{1}{2}$ for all i, because there is always the transition to (t,i) from where the system terminates with probability $\frac{1}{2}$. Consequently, $\mathrm{Val}(\mathit{Term}, (r,i)) \geq \frac{1}{2}$ for all i as well.

Thus, for a fixed i, either $\mathrm{Val}(\mathit{Term}, (s,i)) = \frac{1}{2}$. or $\mathrm{Val}(\mathit{Term}, (s,i)) > \frac{1}{2}$ In the first case, taking the transition $(s,i) \to (r,i)$ is still value-optimal, i.e., $\mathrm{Val}(\mathit{Term}, (s,i)) = \frac{1}{2} \leq \mathrm{Val}(\mathit{Term}, (r,i))$. In the second case the transition $(s,i) \to (t,i)$ is not value-optimal, and thus the transition $(s,i) \to (r,i)$ has to be value optimal.

Thus we know that $(s,i) \to (r,i)$ always preserves the termination value, and we may unfold two steps of the Bellman-style equations satisfied by the value to obtain

$$\mathrm{Val}(\mathit{Term},(s,i)) = \frac{1}{3}\cdot \mathrm{Val}(\mathit{Term},(s,i-1)) + \frac{2}{3}\cdot \mathrm{Val}(\mathit{Term},(s,i+1)).$$

Given the initial conditions for $i = 0,1$, we obtain $\mathrm{Val}(\mathit{Term},(s,i)) = \frac{2^i+1}{2^{i+1}}$ as a unique solution of this recurrence. □

Thus for all $n \geq 1$ we have $\mathrm{Val}(\mathit{Term},(s,n)) = 2^{-(n+1)}\cdot(2^n+1)$, and also, obviously, $\mathrm{Val}(\mathit{Term},(t,n)) = 1/2$. As a conclusion, $\mathrm{Val}(\mathit{Term},(s,n)) > \mathrm{Val}(\mathit{Term},(t,n))$. Thus no termination-optimal strategy may choose a transition generated by the rule $(s,0,t)$. On the other hand, as shown in the first Claim, without the rule $(s,0,t)$ we would have $\mathrm{Val}(\mathit{Term},(s,n)) = 2^{-n} < 2^{-(n+1)}\cdot(2^n+1)$. Consequently, there are no termination-optimal strategies in (s,n).

A.2. Reduction to Rising OC-MDPs

Recall from Definition 3.4 that a pure counterless strategy, σ, is called *idling* if there is a state $q \in Q$, such that $\mathbb{P}^{\sigma}_{(q,0)}\Big(\exists i > 0 : \mathrm{State}^{(i)} = q\Big) = 1$ and for all $i \geq 0$: $\mathbb{P}^{\sigma}_{(q,0)}\Big(\mathrm{State}^{(i)} = q \implies \mathrm{C}^{(i)} = 0\Big) = 1$. Also recall that a maximizing OC-MDP $\mathcal{A}$ is *rising* if there is no idling strategy for $\mathcal{A}$ and, moreover, $\mathrm{Val}(\mathit{LimInf}(=-\infty), q) < 1$ for all states q of $\mathcal{A}$. Before we start proving Lemma 3.10 let us prove an auxiliary result.

Lemma A.2. *Let w be a finite path of length n such that for all $\omega \in \mathit{Run}(w)$:*

- *$C^{(i)}(\omega) > C^{(0)}(\omega)$ for all $i < n$, and*
- *if $0 \leq t < t' \leq n$ and $\mathit{State}^{(t)}(\omega) = \mathit{State}^{(t')}(\omega)$ then $C^{(t)}(\omega) > C^{(t')}(\omega)$.*

Then $n \leq |Q|^2$ and $\max_{0\leq i\leq n} C^{(i)}(\omega) - C^{(0)}(\omega) \leq |Q|$ for all $\omega \in \mathit{Run}(w)$.

Proof. From the fact that the maximal positive counter change is +1 and the second property of w, we have that $\mathrm{C}^{(i)}(\omega) - \mathrm{C}^{(0)}(\omega) < |Q|$ for all $i < n$. Again by the second property, every control state is thus visited at most $|Q|$ times before the counter drops below $\mathrm{C}^{(0)}$. By the first property we now have $n \leq |Q|^2$. □

Proof of Lemma 3.10. We first intuitively explain the idea for the construction of $\bar{\mathcal{A}}$: Using some added information in the control states, the OC-MDP will offer the following possibilities as long as a counterless strategy is chosen: either the

chosen counterless strategy somehow makes sure that after any state s is reached with positive probability the counter will thereafter either be decreased by at least one in at most $|Q|^2$ steps with positive probability, or else after s is reached the play will be forced to enter the "trap" state with positive probability. The "trap" state is an extra absorbing state that keeps increasing the counter value forever thereafter.

This, firstly, ensures that given a OC-MDP, $\mathcal{A}$, the newly constructed OC-MDP, $\bar{\mathcal{A}}$ that is derived from it has no idling strategies. Secondly, the construction ensures the following: for every state q of the original OC-MDP, $\mathcal{A}$, there is a corresponding state $\bar{q}$ of the newly constructed OC-MDP, $\bar{\mathcal{A}}$, such that the optimal termination probability starting at configuration (q, i) in $\mathcal{A}$ is equal to the optimal terimation probability starting at configuration $(\bar{q}, i)$ in $\bar{\mathcal{A}}$.

In more detail, the set $\bar{Q}$ of control states of $\bar{\mathcal{A}}$ will consist of one special state "trap", and of multiple copies of states Q enhanced with two counters. These enhanced states are 3- and 5-tuples of the form $\langle q, n, m\rangle$, $[q, n, m, k, r]$, where $q \in Q$, $(q, k, r) \in \Delta$, $0 \le m \le |Q|^2 + 1$ is a counter measuring the number of steps until it exceeds $|Q|^2 + 1$, and $0 \le n \le |Q| + 1$ is a counter measuring the difference of the current counter value minus the initial one, until it drops below 0 or goes above $|Q|$.

The triples and 5-tuples alternate in the transitions of $\bar{\mathcal{A}}$. First comes a triple, $\langle q, n, m\rangle$, indicating the current configuration of the simulation of a play in $\mathcal{A}$. Then the player has to commit to an outgoing rule, (q, k, r), used on the short path (which it claims exists) which decreases the counter. This results in entering $[q, n, m, k, r]$. If $q \in Q_\top$ then the play must move in the next step to $\langle r, n+k, m+1\rangle$. If $q \in Q_P$ then all outgoing rules for q are used in the next step of the simulation, but the counters m and n are reset to 0 for all steps following rules other than (q, k, r). Thus the next possible triples to visit are $\langle r, n + k, m + 1\rangle$, corresponding to rule (q, k, r), and states $\langle r', 0, 0\rangle$ corresponding to rules (q, k', r'), where $(k', r') \neq (k, r)$. The state in $\bar{\mathcal{A}}$ corresponding to a q in $\mathcal{A}$ is $\langle q, 0, 0\rangle$. Starting at state $\langle q, 0, 0\rangle$ the states along a run in $\bar{\mathcal{A}}$ keep track of the number of steps and the change in counter value, and if the number of steps "overflows" before the counter decreases to -1, this indicates that the path selected by the player is not a short decreasing path, and the simulation is aborted by transiting to the trap state, which results in an incrementing self-loop. Otherwise, if within a short number of m of steps we reach a state $[q', 0, m, q'', k]$, where $m \le |Q|^2$, and the next transition decreases the counter (i.e., $k = -1$) then the two "internal counters" are reset to 0 and we start all over again.

We now give a formal definition of $\bar{\mathcal{A}}$, which is an adaptation of a similar

contruction given in [3], where it appeared as $\mathcal{D}'$. The set of control states of $\bar{\mathcal{A}}$ is

$$\bar{Q} = \{\mathsf{trap}\} \cup \{\langle q, n, m\rangle \mid q \in Q, 0 \leq m \leq |Q|^2 + 1, 0 \leq n \leq |Q| + 1\} \\ \{[q, n, m, k, r] \mid (q, k, r) \in \Delta, 0 \leq m \leq |Q|^2 + 1, 0 \leq n \leq |Q| + 1\}.$$

The stochastic states are $\bar{Q}_P := \{\mathsf{trap}\} \cup \{[q, n, m, k, r] \in \bar{Q}\}$. The rules, $\bar{\Delta}$, is the smallest set containing

$$\{([q, n, |Q|^2 + 1, k, r], 1, \mathsf{trap}) \mid (q, k, r) \in \Delta, 0 \leq n \leq |Q| + 1\}$$
$$\cup \{([q, |Q| + 1, m, k, r], 1, \mathsf{trap}) \mid (q, k, r) \in \Delta, 0 \leq m \leq |Q|^2\}$$
$$\cup \{(\langle q, n, m\rangle, 0, [q, n, m, k, r]) \mid (q, k, r) \in \Delta, 0 \leq n \leq |Q| + 1, 0 \leq m \leq |Q|^2 + 1\}$$
$$\cup \{([q, n, m, k, r], k, \langle r, n + k, m + 1\rangle) \mid (q, k, r) \in \Delta, 0 \leq n \leq |Q|, n + k \geq 0, 0 \leq m \leq |Q|^2\}$$
$$\cup \{([q, n, m, k, r], k, \langle r, 0, 0\rangle) \mid (q, k, r) \in \Delta, n + k = -1, 0 < m \leq |Q|^2\}$$
$$\cup \{([q, n, m, k, r], k', \langle r', 0, 0\rangle) \mid (q, k', r') \in \Delta, q \in Q_P, r' \neq r, 0 \leq n \leq |Q|, 0 \leq m \leq |Q|^2\}$$
$$\cup \{(\mathsf{trap}, 1, \mathsf{trap})\},$$

and also containing the rule $(\bar{q}, 1, \bar{q})$ for each state not having an outgoing rule in the set above. Finally, $\bar{P}$ is derived from P as follows: for all $\bar{q} \in \bar{Q}_P$ which only have one outgoing rule the probability of such rule is 1. Otherwise we know $\bar{q} = [q, n, m, k, r]$, $q \in Q_P$ and can set $\bar{P}(([q, n, m, k, r], k', \langle r', n', m'\rangle)) = P((q, k', r'))$ for each $([q, n, m, k, r], k', \langle r', n', m'\rangle) \in \Delta$.

Clearly, $\|\bar{\mathcal{A}}\| \in O(\|\mathcal{A}\|^4)$. For f we choose the function $f(q) = \langle q, 0, 0\rangle$. The remaining three properties of $\bar{\mathcal{A}}$ are delivered by Lemma A.3, Lemma A.4, and Lemma A.5. □

Lemma A.3. *There are no idling strategies in $\bar{A}$.*

Proof. By contradiction, assume there is a pure counterless idling strategy, σ. From the definition of idling, there is a control state $\bar{q} \in \bar{Q}$ which is almost surely revisited under σ, and upon every revisit, the counter has the same value as at the beginning. For every state $\bar{r}$ visited from $\bar{q}$, i.e., such that $\mathbb{P}^{\sigma}_{(\bar{q},0)}\left(\exists i \geq 0 : \mathrm{State}^{(i)} = \bar{r}\right) > 0$, we define the set of possible counter values seen at a visit from $(\bar{q}, 0)$ to $\bar{r}$ as $C_{\bar{r}} := \{c \in \mathbb{Z} \mid \exists i \geq 0 : \mathbb{P}^{\sigma}_{(\bar{q},0)}\left(\mathrm{State}^{(i)} = \bar{r} \wedge \mathrm{C}^{(i)} = c\right) > 0\}$. First we observe that $|C_{\bar{r}}| = 1$ for all such $\bar{r}$. Indeed, it has obviously at least one element. On the other hand, if $c, d \in C_{\bar{r}}$, $c \neq d$, then $\mathbb{P}^{\sigma}_{(\bar{q},0)}\left(\exists i > 0 : \mathrm{State}^{(i)} = \bar{q} \wedge \mathrm{C}^{(i)} = c - d\right) > 0$ which contradicts our choice of $\bar{q}$

because $c - d \neq 0$. From now on we denote by $c_{\bar{r}}$ the only number such that $C_{\bar{r}} = \{c_{\bar{r}}\}$.

Now we choose $\bar{r}$ so that $c_{\bar{r}}$ is minimal. Observe that $\mathbb{P}^{\sigma}_{(\bar{r},0)}\Big(\mathrm{C}^{(i)} \geq 0\Big) = 1$ for all $i \geq 0$, otherwise there is a state, $\bar{t}$, reachable under σ from $\bar{r}$ such that $c_{\bar{t}} < c_{\bar{r}}$. But this means that a run from $\bar{r}$ under σ visits trap almost surely. It is easy to see that this implies that a run from $\bar{q}$ visits trap with a positive probability.[7] This contradicts σ being idling and $\bar{q}$ being the witnessing state for idling. □

Before we prove the second important property of $\bar{\mathcal{A}}$ we promised, we note that although technically it is not true that $Q \subseteq \bar{Q}$, we may insert Q into $\bar{Q}$ by mapping q to $\langle q, 0, 0\rangle$.

Lemma A.4. $\mathrm{Val}(\mathit{Term}, (q, i)) = \mathrm{Val}(\mathit{Term}, (\langle q, 0, 0\rangle, i))$ *for all* $q \in Q$ *and* $i \geq 0$.

Proof. The inequality $\mathrm{Val}(\mathit{Term}, (q, i)) \geq \mathrm{Val}(\mathit{Term}, (\langle q, 0, 0\rangle, i))$ is easy, because a strategy in $\mathcal{A}$ can simulate a strategy in $\bar{\mathcal{A}}$ (by "projecting" it onto states of $\mathcal{A}$), except for the case when the run in $\mathcal{A}$ reaches trap. But after reaching trap no run terminates, so the simulation in $\mathcal{A}$ may continue arbitrarily without producing a lower probability of termination.

To prove $\mathrm{Val}(\mathit{Term}, (q, i)) \leq \mathrm{Val}(\mathit{Term}, (\langle q, 0, 0\rangle, i))$, we need to show that there are ε-optimal strategies for $\mathcal{A}$, for arbitrarily small $\epsilon > 0$, which can be simulated in $\bar{\mathcal{A}}$ while keeping the termination probability ε-close to the original optimal termination value in $\mathcal{A}$. In the simulation we will use a natural correspondence of paths in $\bar{\mathcal{A}}$ to paths in $\mathcal{A}$, given by dropping the odd steps and all additional information. As an example, the path $(\langle q, 0, 0\rangle, 0) \to ([q, 0, 0, +1, r], 0) \to (\langle r, 1, 1\rangle, 1)$ corresponds to $(q, 0) \to (r, 1)$.

In the proof, we give for every $\varepsilon > 0$ a pure strategy σ_ε, and a measurable set of runs, $T_\varepsilon \subseteq \mathit{Term}$ in $\mathcal{A}$, such that for all $q \in Q$ and $i \geq 0$:

- $\mathbb{P}^{\sigma_\varepsilon}_{(q,i)}(T_\varepsilon) \geq \mathrm{Val}(\mathit{Term}, (q, i)) - \varepsilon$, and
- for all finite paths u, $\mathit{len}(u) = n$, such that $\mathbb{P}^{\sigma_\varepsilon}_{(q,i)}(\mathit{Run}(u) \cap T_\varepsilon) > 0$, there is some k, $n < k \leq n + |Q|^2 + 1$ for which

$$\mathbb{P}^{\sigma_\varepsilon}_{(q,i)}\Big(\mathrm{C}^{(k)} < \mathrm{C}^{(n)} \wedge \forall j, n < j < k : 0 \leq \mathrm{C}^{(j)} - \mathrm{C}^{(n)} \leq |Q| \mid \mathit{Run}(u)\Big) > 0. \qquad (6)$$

Once we have proved the above, we can simulate the strategy σ_ε in $\bar{\mathcal{A}}$.

[7] Actually this probability is again 1, but we only need to know that it is positive.

Let us define the simulation in detail. Let $\bar{u}$ be a path from configuration $(\langle q,0,0\rangle, i)$ in $\bar{\mathcal{A}}$, ending in some configuration $(\langle r,0,0\rangle, j)$, and u the corresponding path in $\mathcal{A}$, ending in (r,j). Let $n = len(u)$. If $\mathbb{P}^{\sigma_\varepsilon}_{(q,i)}(Run(u) \cap T_\varepsilon) = 0$, then the rest of the simulating strategy after intial path $\bar{u}$ can be defined arbitrarily. For later reference we call $\bar{u}$ and all its extensions *dead* in this case. Otherwise let w be some extension of u witnessing (6), i.e., w of length $k \leq n + |Q|^2 + 1$ with a prefix u, such that $\mathbb{P}^{\sigma_\varepsilon}_{(q,i)}(Run(w)) > 0$ and $\mathbb{P}^{\sigma_\varepsilon}_{(q,i)}\Big(\mathrm{C}^{(k)} < \mathrm{C}^{(n)} \wedge \forall j, n < j < k : 0 \leq \mathrm{C}^{(j)} - \mathrm{C}^{(n)} \leq |Q| \mid Run(w)\Big) = 1$.

We fix a unique choice of such a w (which depends on u), and we define the simulating strategy in $\bar{\mathcal{A}}$ for all histories $\bar{v}$ such that the path v in $\mathcal{A}$ which corresponds to $\bar{v}$ is an extension of u and a proper prefix of w. The definition is by induction on the length of $\bar{v}$. Such a history $\bar{v}$ ends in some $\langle s,h,m\rangle$, where $0 \leq m < |Q|^2 + 1$ and $h \geq 0$. Let (s,c) be the last configuration on v and $(s,c) \to (s',c')$ the next step in w after completing v. Then the rule chosen with probability 1 by the simulating strategy in $\bar{A}$ for the history $\bar{v}$ is $(\langle s,h,m\rangle, 0, [s,h,m,c'-c,s'])$. Observe that due to our choice of w, the control state visited in the simulation after completing $\bar{v}$ and visiting $[s,h,m,c'-c,s']$ is either (a) $\langle s', h+c'-c, m+1\rangle$, where $m+1 < |Q|^2 + 1$ and $h + c' - c \geq 0$, or else (b) a state $\langle s'', 0, 0\rangle$, for some state $s'' \in Q$. In the former case (a) we continue with a new $\bar{v}$ as above. In the latter case (b), we are again back in a state of the form $\langle r,0,0\rangle$, and thus we need to find a new extension w' (unless now we are in a dead history) and start the process all over again. Because every history in the simulation is either dead, or ends in some $(\langle r,0,0\rangle, j)$, or is some short extension $\bar{v}$ of such a history which is not dead and ends in some state $\langle r,0,0\rangle$, as above, we have now defined the simulating strategy for every history in $\bar{\mathcal{A}}$.

Moreover, consider a path $\bar{u}$ in $\bar{\mathcal{A}}$, which is not dead. Because we could not possibly hit trap in $\bar{\mathcal{A}}$ before reaching a dead history, and because σ_ε is pure, the probability of $Run(\bar{u})$ in the simulation is $\mathbb{P}^{\sigma_\varepsilon}_{(q,i)}(Run(u))$, where u is the path corresponding to $\bar{u}$ in $\mathcal{A}$. As a consequence, once we prove the existence of σ_ε and validity of its properties, we have proven that the termination value in the simulation is at least $\mathrm{Val}(Term,(q,i)) - \varepsilon$. Because $\varepsilon > 0$ can be chosen arbitrarily, this proves $\mathrm{Val}(Term,(q,i)) \leq \mathrm{Val}(Term,(\langle q,0,0\rangle, i))$. In the rest of the proof we show how to construct such a strategy σ_ε in $\mathcal{A}$ for every $\epsilon > 0$.

Let $t \geq 0$. By $Term^{\leq t}$ we denote the event that $\mathrm{C}^{(t')} = 0$ for some $t' \leq t$. By standard facts (see, e.g., [16, Theorem 4.3.3]), for all t there is a pure strategy τ_t optimal for $Term^{\leq t}$, i.e., such that for all $q \in Q$, $i \geq 0$: $\mathbb{P}^{\tau_t}_{(q,i)}(Term^{\leq t}) = \mathrm{Val}(Term^{\leq t},(q,i))$. Also, easily $\lim_{t\to\infty} \mathrm{Val}(Term^{\leq t},(q,i)) = \mathrm{Val}(Term,(q,i))$, thus

for all $\varepsilon > 0$ there is t_ε such that $\mathrm{Val}(\mathit{Term}^{\leq t_\varepsilon}, (q,i)) \geq \mathrm{Val}(\mathit{Term}, (q,i)) - \varepsilon$. We set $T_\varepsilon := \mathit{Term}^{\leq t_\varepsilon}$.

Let us fix an $\varepsilon > 0$, and consider the corresponding t_ε. We now define σ_ε. Let u be a path in $\mathcal{A}$, of length $n < t_\varepsilon$, ending in a configuration (r, j). Pick the least $t \leq t_\varepsilon - n$ such that $\mathrm{Val}(\mathit{Term}^{\leq t}, (q,i)) = \mathrm{Val}(\mathit{Term}^{\leq t_\varepsilon - n}, (q,i))$. Then $\sigma_\varepsilon(u) = \tau_t((r,j))$. For u where $\mathit{len}(u) \geq t_\varepsilon$ we define $\sigma_\varepsilon(u)$ arbitrarily. Due to the Bellman-equation characterization of optimality for finite-horizon objectives, given, e.g., in [16, Section 4.3], we obtain that for all configurations (q,i): $\mathbb{P}^{\sigma_\varepsilon}_{(q,i)}(\mathit{Term}) \geq \mathbb{P}^{\sigma_\varepsilon}_{(q,i)}(\mathit{Term}^{\leq t_\varepsilon}) = \mathrm{Val}(\mathit{Term}^{\leq t_\varepsilon}, (q,i)) \geq \mathrm{Val}(\mathit{Term}, (q,i)) - \varepsilon$, as required.

For $k \geq 0$, let E_k be the event that there are times t, t', $k \leq t < t'$, such that $\mathrm{State}^{(t)} = \mathrm{State}^{(t')}$ and $0 < \mathrm{C}^{(t)} \leq \mathrm{C}^{(t')}$.

Claim 4. *Let $k \geq 0$, and (q,i) be a configuration. Further, let u be an arbitrary path such that $\mathit{len}(u) = k$, $\mathbb{P}^{\sigma_\varepsilon}_{(q,i)}(\mathit{Run}(u)) > 0$ and $\mathbb{P}^{\sigma_\varepsilon}_{(q,i)}(\mathit{Term}^{\leq t_\varepsilon} \mid \mathit{Run}(u)) > 0$. Then $\mathbb{P}^{\sigma_\varepsilon}_{(q,i)}(E_k \mid \mathit{Run}(u) \cap \mathit{Term}^{\leq t_\varepsilon}) < 1$.*

Proof. By contradiction. For $k \geq t_\varepsilon$ the statement is obvious. Fix some k, $0 \leq k < t_\varepsilon$ and (q,i). Assume that $\mathbb{P}^{\sigma_\varepsilon}_{(q,i)}(E_k \mid \mathit{Run}(u) \cap \mathit{Term}^{\leq t_\varepsilon}) = 1$. Let w be an arbitrary extension of u such that $\mathit{len}(w) = t_\varepsilon$, $\mathit{Run}(w) \subseteq \mathit{Term}^{\leq t_\varepsilon}$, and $\mathbb{P}^{\sigma_\varepsilon}_{(q,i)}(\mathit{Run}(w)) > 0$. Then clearly $\mathit{Run}(w) \subseteq E_k$. This means that there are times t, t', $k \leq t < t' \leq t_\varepsilon$ such that $\mathbb{P}^{\sigma_\varepsilon}_{(q,i)}\Big(\mathrm{State}^{(t)} = \mathrm{State}^{(t')} \mid \mathit{Run}(w)\Big) = 1$, and $\mathbb{P}^{\sigma_\varepsilon}_{(q,i)}\Big(0 < \mathrm{C}^{(t)} \leq \mathrm{C}^{(t')} \mid \mathit{Run}(w)\Big) = 1$. Consider the prefixes $\bar{w}$, $\bar{w}'$ of w of lengths t and t', respectively. There is some state $r \in Q$, and counter values $0 < j \leq j'$ such that $\bar{w}$ ends in (r,j), and $\bar{w}'$ ends in (r,j'). By the construction of σ_ε, there are $h \leq t_\varepsilon - t$ and $h' \leq t_\varepsilon - t'$ such that $h' < h$, $\mathrm{Val}(\mathit{Term}^{\leq h}, (r,j)) > \mathrm{Val}(\mathit{Term}^{\leq h-1}, (r,j)) \geq \mathrm{Val}(\mathit{Term}^{\leq h'}, (r,j'))$, and $\mathbb{P}^{\sigma_\varepsilon}_{(q,i)}(\mathit{Term}^{\leq t_\varepsilon} \mid \mathit{Run}(\bar{w})) = \mathrm{Val}(\mathit{Term}^{\leq h}, (r,j))$, $\mathbb{P}^{\sigma_\varepsilon}_{(q,i)}(\mathit{Term}^{\leq t_\varepsilon} \mid \mathit{Run}(\bar{w}')) = \mathrm{Val}(\mathit{Term}^{\leq h'}, (r,j'))$. In other words, on every extension of u which eventually satisfies $\mathit{Term}^{\leq t_\varepsilon}$ there is a moment where the probability of $\mathit{Term}^{\leq t_\varepsilon}$, conditionally on the current history, sharply decreases. This is in contradiction with the fact that σ_ε is optimal wrt. $\mathit{Term}^{\leq t_\varepsilon}$, and thus satisfies the Bellman optimality equations (cf. [16, Section 4.3]).

□

Let us fix an arbitrary path u, $\mathit{len}(u) = n$, such that $\mathbb{P}^{\sigma_\varepsilon}_{(q,i)}(\mathit{Run}(u) \cap \mathit{Term}^{\leq t_\varepsilon}) > 0$. We apply Claim 4, and obtain a witnessing extension, w, $\mathit{len}(w) = m$, of u so that $\mathbb{P}^{\sigma_\varepsilon}_{(q,i)}(\mathit{Run}(w)) > 0$, $\mathbb{P}^{\sigma_\varepsilon}_{(q,i)}\Big(\mathrm{C}^{(m)} = 0 \mid \mathit{Run}(w)\Big) = 1$, and $\mathbb{P}^{\sigma_\varepsilon}_{(q,i)}(E_n \mid \mathit{Run}(w)) = 0$. By Lemma A.2 this implies that there is some k, $n < k \leq n + |Q|^2 + 1$ such that (6) is satisfied. Thus we proved all the required properties of σ_ε and $T_\varepsilon = \mathit{Term}^{\leq t_\varepsilon}$, and the proof is done. □

As a consequence, we obtain the last promised property of $\bar{\mathcal{A}}$.

Lemma A.5. *If* $\mathrm{Val}(\mathit{LimInf}(=-\infty), q) < 1$ *for all* $q \in Q$ *in* $\mathcal{A}$*, then* $\bar{\mathcal{A}}$ *is rising.*

Proof. By Lemma A.3 there are no idling strategies in $\bar{\mathcal{A}}$. It remains to prove that $\mathrm{Val}(\mathit{LimInf}(=-\infty), \bar{q}) < 1$ for all $q \in \bar{Q}$ in $\bar{\mathcal{A}}$. First we prove it for $\bar{q}$ of the form $\langle q, 0, 0\rangle$. If $\mathrm{Val}(\mathit{LimInf}(=-\infty), q) < 1$ then there is $i \geq 0$ such that $\mathrm{Val}(\mathit{Term}, (q, i)) < 1$, by, e.g., Lemma 14 of [2]. By Lemma A.4 we thus have $\mathrm{Val}(\mathit{Term}, (\bar{q}, i)) < 1$, and, thus again by Lemma 14 of [2], $\mathrm{Val}(\mathit{LimInf}(=-\infty), \bar{q}) < 1$. If $\bar{q} = \mathsf{trap}$ we are done immediately, as $\mathrm{Val}(\mathit{LimInf}(=-\infty), \mathsf{trap}) = 0$. Finally, in all remaining cases of $\bar{q}$ the play will almost surely reach some states from $\{\mathsf{trap}\} \cup \{\langle q, 0, 0\rangle \mid q \in Q\}$. Because $\mathit{LimInf}(=-\infty)$ is prefix independent, $\mathrm{Val}(\mathit{LimInf}(=-\infty), \bar{q}) < 1$ also in this case, and the proof is finished. □

A.3. Bounds on N

Here we derive an exponential upper bound on the value N, introduced in Section 3. Recall that, given a OC-SSG, $\mathcal{A} = (Q, \Delta, P)$, and an $\varepsilon > 0$, we want N to satisfy:

$$\mathrm{Val}(\mathit{Term}, (q, i)) - \mathrm{Val}(\mathit{LimInf}(=-\infty), q) \leq \varepsilon \quad \text{for all } q \in Q \text{ and } i \geq N.$$

By results of Section 3.2, it suffices to consider only the case when $\mathcal{A}$ is a maximizing OC-MDP. From Section 3.1 we know that $N := \max\{h, \lceil \log_c(\varepsilon \cdot (1-c))\rceil\}$, where $c = \exp\left(\frac{-\bar{x}^2}{2\cdot(\bar{z}_{\max}+\bar{x}+1)}\right) < 1$ and $h = \lceil \bar{z}_{\max}\rceil$, and $\bar{x}$ and $\bar{z}_{\max}$ are solutions to a linear program with coefficients polynomial in $\|\mathcal{A}\|$. Thus there is a positive polynomial, p, such that $c \leq \mathrm{e}^{-\mathrm{e}^{-p(\|\mathcal{A}\|)}}$ and $h \leq \mathrm{e}^{p(\|\mathcal{A}\|)}$. If $N \leq h$ we have clearly that it is exponentially bounded in $\|\mathcal{A}\|$. Otherwise

$$N \leq f_\varepsilon(c) := \frac{\ln(\varepsilon) + \ln(1-c)}{\ln(c)}.$$

Observe that $f_\varepsilon(c)$ is growing with $c \to 1^-$ and fixed ε, because $\frac{c^{f_\varepsilon(c)}}{1-c} = \varepsilon$ and $\frac{c^i}{1-c}$ grows with $c \to 1^-$ and fixed i. Thus

$$\begin{aligned} N &\leq f_\varepsilon(c) \leq f_\varepsilon(\mathrm{e}^{-\mathrm{e}^{-p(\|\mathcal{A}\|)}}) \\ &= \frac{\ln(\varepsilon) + \ln(1-\mathrm{e}^{-\mathrm{e}^{-p(\|\mathcal{A}\|)}})}{-\mathrm{e}^{-p(\|\mathcal{A}\|)}} = \mathrm{e}^{p(\|\mathcal{A}\|)} \cdot \ln(1/\varepsilon) - \ln(1-\mathrm{e}^{-\mathrm{e}^{-p(\|\mathcal{A}\|)}}) \cdot \mathrm{e}^{p(\|\mathcal{A}\|)}. \quad (7) \end{aligned}$$

Before we prove that this is indeed an exponential bound on N, let us prove two auxiliary claims.

Claim 5. *For all $n \geq 0$ the following inequality holds:*

$$\mathrm{e}^{-1} - \mathrm{e}^{-1-\mathrm{e}^{-n}} \leq 1 - \mathrm{e}^{-\mathrm{e}^{-(n+1)}}. \tag{8}$$

Proof. We set $d(n) \coloneqq \mathrm{e}^{-\mathrm{e}^{-(n+1)}} - \mathrm{e}^{-1-\mathrm{e}^{-n}}$. The inequality (8) is equivalent to $d(n) \leq 1 - \mathrm{e}^{-1}$. Because $\lim_{n\to\infty} d(n) = 1 - \mathrm{e}^{-1}$, it suffices to prove that $d(n)$ is increasing: Observe that

$$d(n+1) - d(n) = (\mathrm{e}^{-\mathrm{e}^{-(n+2)}} - \mathrm{e}^{-\mathrm{e}^{-(n+1)}}) - \mathrm{e}^{-1} \cdot (\mathrm{e}^{-\mathrm{e}^{-(n+1)}} - \mathrm{e}^{-\mathrm{e}^{-n}}). \tag{9}$$

Also, because the exponential function e^x is increasing and has increasing derivation $\mathrm{e}^x \geq 0$, we know that

$$\frac{\mathrm{e}^a - \mathrm{e}^b}{\mathrm{e}^b - \mathrm{e}^c} \geq \frac{a-b}{b-c} \quad \text{for all } a > b > c.$$

In particular, setting $a = -\mathrm{e}^{-(n+2)}$, $b = -\mathrm{e}^{-(n+1)}$, and $c = -\mathrm{e}^{-n}$ yields

$$\frac{\mathrm{e}^{-\mathrm{e}^{-(n+2)}} - \mathrm{e}^{-\mathrm{e}^{-(n+1)}}}{\mathrm{e}^{-\mathrm{e}^{-(n+1)}} - \mathrm{e}^{-\mathrm{e}^{-n}}} \geq \mathrm{e}^{-1}.$$

By (9), this implies $d(n+1) \geq d(n)$ as required. □

Claim 6. *For all $n \geq 0$ the following inequality holds:*

$$n + 1 \geq -\ln(1 - \mathrm{e}^{-\mathrm{e}^{-n}}). \tag{10}$$

Proof. By induction. A direct computation for $n = 0$ shows $-\ln(1 - \mathrm{e}^{-\mathrm{e}^{-0}}) = -\ln(1 - \mathrm{e}^{-1}) \leq 0.46 < 1$. Consider now $n = k + 1$ for some $k \geq 0$. Using (8) and the inductive hypothesis, we obtain

$$(k+1) + 1 \geq -\ln(1 - \mathrm{e}^{-\mathrm{e}^{-k}}) + 1 = -\ln(\mathrm{e}^{-1} - \mathrm{e}^{-1-\mathrm{e}^{-k}}) \geq -\ln(1 - \mathrm{e}^{-\mathrm{e}^{-(k+1)}}).$$

□

Finally, using (10) in (7) we get $N \leq \mathrm{e}^{p(\|\mathcal{A}\|)} \cdot \ln(1/\varepsilon) + (1 + p(\|\mathcal{A}\|)) \cdot \mathrm{e}^{p(\|\mathcal{A}\|)}$.