# Minimizing Expected Termination Time in One-Counter Markov Decision Processes

Tomáš Brázdil[1]$^\star$, Antonín Kučera[1]$^\star$, Petr Novotný[1]$^\star$, and Dominik Wojtczak[2]$^\star$

[1] Faculty of Informatics, Masaryk University
{xbrazdil,kucera,xnovot18}@fi.muni.cz
[2] Department of Computer Science, University of Liverpool
d.wojtczak@liv.ac.uk

**Abstract.** We consider the problem of computing the value and an optimal strategy for minimizing the expected termination time in one-counter Markov decision processes. Since the value may be irrational and an optimal strategy may be rather complicated, we concentrate on the problems of approximating the value up to a given error $\varepsilon > 0$ and computing a finite representation of an $\varepsilon$-optimal strategy. We show that these problems are solvable in exponential time for a given configuration, and we also show that they are computationally hard in the sense that a polynomial-time approximation algorithm cannot exist unless P=NP.

## 1 Introduction

In recent years, a lot of research work has been devoted to the study of stochastic extensions of various automata-theoretic models such as pushdown automata, Petri nets, lossy channel systems, and many others. In this paper we study the class of *one-counter Markov decision processes (OC-MDPs)*, which are infinite-state MDPs [19, 13] generated by finite-state automata operating over a single unbounded counter. Intuitively, an OC-MDP is specified by a finite directed graph $\mathcal{A}$ where the nodes are control states and the edges correspond to transitions between control states. Each control state is either stochastic or non-deterministic, which means that the next edge is chosen either randomly (according to a fixed probability distribution over the outgoing edges) or by a controller. Further, each edge either increments, decrements, or leaves unchanged the current counter value. A *configuration* $q(i)$ of an OC-MDP $\mathcal{A}$ is given by the current control state $q$ and the current counter value $i$ (for technical convenience, we also allow negative counter values, although we are only interested in runs where the counter stays non-negative). The outgoing transitions of $q(i)$ are determined by the edges of $\mathcal{A}$ in the natural way.

Previous works on OC-MDPs [4, 2, 3] considered mainly the objective of *maximizing/minimizing termination probability*. We say that a run initiated in a configuration $q(i)$ *terminates* if it visits a configuration with zero counter. The goal of the controller is to play so that the probability of all terminating runs is maximized (or minimized).

$^\star$ Tomáš Brázdil is supported by the Czech Science Foundation, grant No. P202/12/P612. Antonín Kučera and Petr Novotný are supported by the Czech Science Foundation, grant No. P202/10/1469. Dominik Wojtczak is supported by EPSRC grant EP/G050112/2.

In this paper, we study a related objective of *minimizing the expected termination time*. Formally, we define a random variable $T$ over the runs of $\mathcal{A}$ such that $T(\omega)$ is equal either to $\infty$ (if the run $\omega$ is non-terminating) or to the number of transitions needed to reach a configuration with zero counter (if $\omega$ is terminating). The goal of the controller is to minimize the expectation $\mathbb{E}(T)$. The *value* of $q(i)$ is the infimum of $\mathbb{E}(T)$ over all strategies. It is easy to see that the controller has a memoryless deterministic strategy which is optimal (i.e., achieves the value) in every configuration. However, since OC-MDPs have infinitely many configurations, this does not imply that an optimal strategy is finitely representable and computable. Further, the value itself can be irrational. Therefore, we concentrate on the problem of *approximating* the value of a given configuration up to a given (absolute or relative) error $\varepsilon > 0$, and computing a strategy which is *$\varepsilon$-optimal* (in both absolute and relative sense). Our main results can be summarized as follows:

- ***The value and optimal strategy can be effectively approximated up to a given relative/absolute error in exponential time.*** More precisely, we show that given an OC-MDP $\mathcal{A}$, a configuration $q(i)$ of $\mathcal{A}$ where $i \geq 0$, and $\varepsilon > 0$, the value of $q(i)$ up to the (relative or absolute) error $\varepsilon$ is computable in time exponential in the encoding size of $\mathcal{A}$, $i$, and $\varepsilon$, where all numerical constants are represented as fractions of binary numbers. Further, there is a history-dependent deterministic strategy $\sigma$ computable in exponential time such that the absolute/relative difference between the value of $q(i)$ and the outcome of $\sigma$ in $q(i)$ is bounded by $\varepsilon$.
- ***The value is not approximable in polynomial time unless P=NP.*** This hardness result holds even if we restrict ourselves to configurations with counter value equal to 1 and to OC-MDPs where every outgoing edge of a stochastic control state has probability $1/2$. The result is valid for absolute as well as relative approximation.

Let us sketch the basic ideas behind these results. The upper bounds are obtained in two steps. In the first step (Section 3.1), we analyze the special case when the underlying graph of $\mathcal{A}$ is strongly connected. We show that minimizing the expected termination time is closely related to minimizing the expected increase of the counter per transition, at least for large counter values. We start by computing the minimal expected increase of the counter per transition (denoted by $\bar{x}$) achievable by the controller, and the associated strategy $\sigma$. This is done by standard linear programming techniques developed for optimizing the long-run average reward in finite-state MDPs (see, e.g., [19]) applied to the underlying finite graph of $\mathcal{A}$. Note that $\sigma$ depends only on the current control state and ignores the current counter value (we say that $\sigma$ is *counterless*). Further, the encoding size of $\bar{x}$ is *polynomial* in the encoding size of $\mathcal{A}$ (which we denote by $||\mathcal{A}||$). Then, we distinguish two cases.

*Case (A),* $\bar{x} \geq 0$. Then the counter does not have a tendency to decrease *regardless* of the controller's strategy, and the expected termination time value is infinite in all configurations $q(i)$ such that $i \geq |Q|$, where $Q$ is the set of control states of $\mathcal{A}$ (see Proposition 5. A). For the finitely many remaining configurations, we can compute the value and optimal strategy precisely by standard methods for finite-state MDPs.

*Case (B),* $\bar{x} < 0$. Then, one intuitively expects that applying the strategy $\sigma$ in an initial configuration $q(i)$ yields the expected termination time about $i/|\bar{x}|$. Actually, this is *almost* correct; we show (Proposition 5. B.2) that this expectation is bounded by $(i + U)/|\bar{x}|$, where $U \geq 0$ is a constant depending only on $\mathcal{A}$ whose size is at most exponential in $||\mathcal{A}||$. Further, we show that an *arbitrary* strategy $\pi$ applied to $q(i)$ yields the expected termination time *at least* $(i - V)/|\bar{x}|$, where $V \geq 0$ is a constant depending only on $\mathcal{A}$ whose size is at most exponential in $||\mathcal{A}||$ (Proposition 5. B.1). In particular, this applies to the *optimal* strategy $\pi^*$ for minimizing the expected termination time. Hence, $\pi^*$ can be more efficient than $\sigma$, but the difference between their outcomes is bounded by a constant which depends only on $\mathcal{A}$ and is at most exponential in $||\mathcal{A}||$. We proceed by computing a sufficiently large $k$ so that the probability of increasing the counter to $i + k$ by a run initiated in $q(i)$ is inevitably (i.e., under any optimal strategy) so small that the controller can safely switch to the strategy $\sigma$ when the counter reaches the value $i + k$. Then, we construct a *finite-state* MDP $\mathcal{M}$ and a reward function $f$ over its transitions such that

- the states are all configurations $p(j)$ where $0 \leq j \leq i + k$;
- all states with counter values less than $i + k$ "inherit" their transitions from $\mathcal{A}$; configurations of the form $p(i + k)$ have only self-loops;
- the self-loops on configurations where the counter equals 0 or $i{+}k$ have zero reward, transitions leading to configurations where the counter equals $i + k$ have reward $(i + k + U)/|\bar{x}|$, and the other transitions have reward 1.

In this finite-state MDP $\mathcal{M}$, we compute an optimal memoryless deterministic strategy $\varrho$ for the total accumulated reward objective specified by $f$. Then, we consider another strategy $\hat{\sigma}$ for $q(i)$ which behaves like $\varrho$ until the point when the counter reaches $i + k$, and from that point on it behaves like $\sigma$. It turns out that the absolute as well as relative difference between the outcome of $\hat{\sigma}$ in $q(i)$ and the value of $q(i)$ is bounded by $\varepsilon$, and hence $\hat{\sigma}$ is the desired $\varepsilon$-optimal strategy.

In the general case when $\mathcal{A}$ is not necessarily strongly connected (see Section 3.2), we have to solve additional difficulties. Intuitively, we split the graph of $\mathcal{A}$ into maximal end components (MECs), where each MEC can be seen as a strongly connected OC-MDP and analyzed by the techniques discussed above. In particular, for every MEC $C$ we compute the associated $\bar{x}_C$ (see above). Then, we consider a strategy which tries to reach a MEC as quickly as possible so that the expected value of the fraction $1/|\bar{x}_C|$ is *minimal*. After reaching a target MEC, the strategy starts to behave as the strategy $\sigma$ discussed above. It turns out that this particular strategy cannot be much worse than the optimal strategy (a proof of this claim requires new observations), and the rest of the argument is similar as in the strongly connected case.

The lower bound, i.e., the result saying that the value cannot be efficiently approximated unless P=NP (see Section 4), seems to be the first result of this kind for OC-MDPs. Here we combine the technique of encoding propositional assignments presented in [17] (see also [15]) with some new gadgets constructed specifically for this proof (let us note that we did not manage to improve the presented lower bound to PSPACE by adapting other known techniques [14, 20, 16]). As a byproduct, our proof also reveals that the optimal strategy for minimizing the expected termination time *cannot* ignore the precise counter value, even if the counter becomes very large. In our example, the (only) optimal strategy is *eventually periodic* in the sense that for a sufficiently large counter value $i$, it is only "$i$ modulo $c$" which matters, where $c$ is a fixed (exponentially large) constant. The question whether there *always* exists an optimal eventually periodic strategy is left open. Another open question is whether our results can be extended to stochastic games over one-counter automata.

Due to space constraints, most proofs are omitted and can be found in [**?**].

**Related Work:** One-counter automata can also be seen as pushdown automata with one letter stack alphabet. Stochastic games and MDPs generated by pushdown automata and stateless pushdown automata (also known as BPA) with termination and reachability objectives have been studied in [11, 12, 5, 6]. To the best of our knowledge, the only prior work on the expected termination time (or, more generally, total accumulated reward) objective for a class of infinite-state MDPs or stochastic games is [9], where this problem is studied for stochastic BPA games. However, the proof techniques of [9] are not directly applicable to one-counter automata.

The termination objective for one-counter MDPs and games has been examined in [4, 2, 3], where it was shown (among other things) that the equilibrium termination probability (i.e., the termination value) can be approximated up to a given precision in exponential time, but no lower bound was provided. In this paper, we build on some of the underlying observations presented in [4, 2, 3]. In particular, we employ the submartingale of [3] to derive certain bounds in Section 3.1.

The games over one-counter automata are also known as "energy games" [7, 8]. Intuitively, the counter is used to model the amount of currently available energy, and the aim of the controller is to optimize the energy consumptions.

Finally, let us note that OC-MDPs can be seen as discrete-time Quasi-Birth-Death Processes (QBDs, see, e.g., [18, 10]) extended with a control. Hence, the theory of one-counter MDPs and games is closely related to queuing theory, where QBDs are considered as a fundamental model.

## 2 Preliminaries

Given a set $A$, we use $|A|$ to denote the cardinality of $A$. We also write $|x|$ to denote the absolute value of a given $x \in \mathbb{R}$, but this should not cause any confusions. The encoding size of a given object $B$ is denoted by $\|B\|$. The set of integers is denoted by $\mathbb{Z}$, and the set of positive integers by $\mathbb{N}$.

We assume familiarity with basic notions of probability theory. In particular, we call a probability distribution $f$ over a discrete set $A$ *positive* if $f(a) > 0$ for all $a \in A$.

**Definition 1 (MDP).** *A* Markov decision process (MDP) *is a tuple* $\mathcal{M} = (S, (S_0, S_1), \rightsquigarrow, Prob)$*, consisting of a countable set of* states $S$ *partitioned into the sets* $S_0$ *and* $S_1$ *of* stochastic *and* non-deterministic *states, respectively. The* edge relation $\rightsquigarrow \subseteq S \times S$ *is total, i.e., for every* $r \in S$ *there is* $s \in S$ *such that* $r \rightsquigarrow s$*. Finally,* $Prob$ *assigns to every* $s \in S_0$ *a positive probability distribution over its outgoing edges.*

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

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

In the rest of this paper we often write $q \xrightarrow{i} r$ to indicate that $(q, i, r) \in \delta$, and $q \xrightarrow{i,x} r$ to indicate that $(q, i, r) \in \delta$, $q$ is stochastic, and $P_q(q, i, r) = x$. Without restrictions, we assume that for each pair $q, r \in Q$ there is at most one $i$ such that $(q, i, r) \in \delta$. The encoding size of $\mathcal{A}$ is denoted by $||\mathcal{A}||$, where all numerical constants are encoded as fractions of binary numbers. The set of all *configurations* is $C := \{q(i) \mid q \in Q, i \in \mathbb{Z}\}$.

To $\mathcal{A}$ we associate an infinite-state MDP $\mathcal{M}_{\mathcal{A}}^{\infty} = (C, (C_0, C_1), \rightsquigarrow, Prob)$, where the partition of $C$ is defined by $q(i) \in C_0$ iff $q \in Q_0$, and similarly for $C_1$. The edges are defined by $q(i) \rightsquigarrow r(j)$ iff $(q, j - i, r) \in \delta$. The probability assignment $Prob$ is derived naturally from $P$ and we write $q(i) \overset{x}{\rightsquigarrow} r(j)$ to indicate that $q(i) \in C_0$, and $Prob(q(i))$ assigns probability $x$ to the edge $q(i) \rightsquigarrow r(j)$.

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

**Strategies and Probability.** Let $\mathcal{M}$ be an MDP. A *history* is a finite path in $\mathcal{M}$, and a *strategy* (or *policy*) is a function assigning to each history ending in a state from $S_1$ a distribution on edges leaving the last state of the history. A strategy $\sigma$ is *pure* (or *deterministic*) if it always assigns 1 to one edge and 0 to the others, and *memoryless* if $\sigma(ws)$ depends just on the last state $s$, for every $w \in S^*$.

Now consider some OC-MDP $\mathcal{A}$. A strategy $\sigma$ over the histories in $\mathcal{M}_{\mathcal{A}}^{\infty}$ is *counterless* if it is memoryless and $\sigma(q(i)) = \sigma(q(j))$ for all $i, j$. Observe that every strategy $\sigma$ for $\mathcal{M}_{\mathcal{A}}$ gives a unique strategy $\sigma'$ for $\mathcal{M}_{\mathcal{A}}^{\infty}$ which just forgets the counter values in the history and plays as $\sigma$. This correspondence is bijective when restricted to memoryless strategies in $\mathcal{M}_{\mathcal{A}}$ and counterless strategies in $\mathcal{M}_{\mathcal{A}}^{\infty}$, and it is used implicitly throughout the paper.

Fixing a strategy $\sigma$ and an initial state $s$, we obtain in a standard way a probability measure $\mathbb{P}_s^{\sigma}(\cdot)$ on the subspace of runs starting in $s$. For MDPs of the form $\mathcal{M}_{\mathcal{A}}^{\infty}$ for some OC-MDP $\mathcal{A}$, we consider two sequences of random variables, $\{C^{(i)}\}_{i \geq 0}$ and $\{S^{(i)}\}_{i \geq 0}$, returning the current counter value and the current control state after completing $i$ transitions.

**Termination Time in OC-MDPs.** Let $\mathcal{A}$ be an OC-MDP. A run $\omega$ in $\mathcal{M}_{\mathcal{A}}^{\infty}$ *terminates* if $\omega(j) = q(0)$ for some $j \geq 0$ and $q \in Q$. The associated *termination time*, denoted by $T(\omega)$, is the least $j$ such that $\omega(j) = q(0)$ for some $q \in Q$. If there is no such $j$, we put $T(\omega) = \infty$, where the symbol $\infty$ denotes the "infinite amount" with the standard conventions, i.e., $c < \infty$ and $\infty + c = \infty + \infty = \infty \cdot d = \infty$ for arbitrary real numbers $c, d$ where $d > 0$.

For every strategy $\sigma$ and a configuration $q(i)$, we use $\mathbb{E}^{\sigma} q(i)$ to denote the expected value of $T$ in the probability space of all runs initiated in $q(i)$ where $\mathbb{P}_{q(i)}^{\sigma}(\cdot)$ is the underlying probability measure. The *value* of a given configuration $q(i)$ is defined by $\mathrm{Val}(q(i)) := \inf_{\sigma} \mathbb{E}^{\sigma} q(i)$. Let $\varepsilon \geq 0$ and $i \geq 1$. We say that a constant $v$

$$\begin{aligned} &\textit{maximize } x, \text{ subject to} \\ &\qquad z_q \le -x + k + z_r && \text{for all } q \in Q_1 \text{ and } (q,k,r) \in \delta, \\ &\qquad z_q \le -x + \textstyle\sum_{(q,k,r)\in\delta} P_q((q,k,r)) \cdot (k + z_r) && \text{for all } q \in Q_0, \end{aligned}$$

**Fig. 1.** The linear program $\mathcal{L}$ over $x$ and $z_q$, where $q \in Q$.

approximates $\mathrm{Val}(q(i))$ up to the absolute or relative error $\varepsilon$ if $|\mathrm{Val}(q(i)) - v| \le \varepsilon$ or $|\mathrm{Val}(q(i)) - v|/\mathrm{Val}(q(i)) \le \varepsilon$, respectively. Note that if $v$ approximates $\mathrm{Val}(q(i))$ up to the absolute error $\varepsilon$, then it also approximates $\mathrm{Val}(q(i))$ up to the relative error $\varepsilon$ because $\mathrm{Val}(q(i)) \ge 1$. A strategy $\sigma$ is (absolutely or relatively) *$\varepsilon$-optimal* if $\mathbb{E}^{\sigma} q(i)$ approximates $\mathrm{Val}(q(i))$ up to the (absolute or relative) error $\varepsilon$. A 0-optimal strategy is called *optimal*.

It is easy to see that there is a memoryless deterministic strategy $\sigma$ in $\mathcal{M}^{\infty}_{\mathcal{A}}$ which is optimal in every configuration of $\mathcal{M}^{\infty}_{\mathcal{A}}$. First, observe that for all $q \in Q_0$, $q' \in Q_1$, and $i \neq 0$ we have that

$$\begin{aligned} \mathrm{Val}(q(i)) &= 1 + \textstyle\sum_{q(i) \overset{x}{\rightsquigarrow} r(j)} x \cdot \mathrm{Val}(r(j)) \\ \mathrm{Val}(q'(i)) &= 1 + \min\{\mathrm{Val}(r(j)) \mid q'(i) \rightsquigarrow r(j)\}. \end{aligned}$$

We put $\sigma(q'(i)) = r(j)$ if $q'(i) \rightsquigarrow r(j)$ and $\mathrm{Val}(q'(i)) = 1 + \mathrm{Val}(r(j))$ (if there are several candidates for $r(j)$, any of them can be chosen). Now we can easily verify that $\sigma$ is indeed optimal in every configuration.

## 3 Upper Bounds

The goal of this section is to prove the following:

**Theorem 3.** *Let $\mathcal{A}$ be an OC-MDP, $q(i)$ a configuration of $\mathcal{A}$ where $i \ge 0$, and $\varepsilon > 0$.*

1. *The problem whether* $\mathrm{Val}(q(i)) = \infty$ *is decidable in polynomial time.*
2. *There is an algorithm that computes a rational number $v$ such that $|\mathrm{Val}(q(i)) - v| \le \varepsilon$, and a strategy $\sigma$ that is absolutely $\varepsilon$-optimal starting in $q(i)$. The algorithm runs in time exponential in $||\mathcal{A}||$ and polynomial in $i$ and $1/\varepsilon$. (Note that $v$ then approximates* $\mathrm{Val}(q(i))$ *also up to the relative error $\varepsilon$, and $\sigma$ is also relatively $\varepsilon$-optimal in $q(i)$).*

For the rest of this section, we fix an OC-MDP $\mathcal{A} = (Q, (Q_0, Q_1), \delta, P)$. First, we prove Theorem 3 under the assumption that $\mathcal{M}_{\mathcal{A}}$ is *strongly connected* (Section 3.1). A generalization to arbitrary OC-MDP is then given in Section 3.2.

### 3.1 Strongly Connected OC-MDP

Let us assume that $\mathcal{M}_{\mathcal{A}}$ is strongly connected, i.e., for all $p, q \in Q$ there is a finite path from $p$ to $q$ in $\mathcal{M}_{\mathcal{A}}$. Consider the linear program of Figure 1. Intuitively, the variable $x$ encodes a lower bound on the long-run trend of the counter value. More precisely, the maximal value of $x$ corresponds to the *minimal* long-run average change in the counter value achievable by any strategy. The program corresponds to the one used for optimizing the long-run average reward in Sections 8.8 and 9.5 of [19], and hence we know it has a solution.

**Lemma 4 ([19]).** *There is a rational solution* $\left(\bar{x}, (\bar{z}_q)_{q\in Q}\right) \in \mathbb{Q}^{|Q|+1}$ *to* $\mathcal{L}$*, and the encoding size*[3] *of the solution is polynomial in* $||\mathcal{A}||$.

Note that $\bar{x} \geq -1$, because for any fixed $x \leq -1$ the program $\mathcal{L}$ trivially has a feasible solution. Further, we put $V := \max_{q\in Q} \bar{z}_q - \min_{q\in Q} \bar{z}_q$. Observe that $V \in \exp\left(||\mathcal{A}||^{O(1)}\right)$ and $V$ is computable in time polynominal in $||\mathcal{A}||$.

**Proposition 5.** *Let* $\left(\bar{x}, (\bar{z}_q)_{q\in Q}\right)$ *be a solution of* $\mathcal{L}$.

*(A) If* $\bar{x} \geq 0$*, then* $\mathrm{Val}(q(i)) = \infty$ *for all* $q \in Q$ *and* $i \geq |Q|$.
*(B) If* $\bar{x} < 0$*, then the following holds:*
  *(B.1) For* every *strategy* $\pi$ *and all* $q \in Q$*,* $i \geq 0$ *we have that* $\mathbb{E}^{\pi} q(i) \;\geq\; (i - V)/|\bar{x}|$.
  *(B.2) There is a counterless strategy* $\sigma$ *and a number* $U \in \exp\left(||\mathcal{A}||^{O(1)}\right)$ *such that for all* $q \in Q$*,* $i \geq 0$ *we have that* $\mathbb{E}^{\sigma} q(i) \leq (i + U)/|\bar{x}|$. *Moreover,* $\sigma$ *and* $U$ *are computable in time polynomial in* $||\mathcal{A}||$.

First, let us realize that Proposition 5 implies Theorem 3. To see this, we consider the cases $\bar{x} \geq 0$ and $\bar{x} < 0$ separately. In both cases, we resort to analyzing a finite-state MDP $\mathcal{G}_K$, where $K$ is a suitable natural number, obtained by restricting $\mathcal{M}^{\infty}_{\mathcal{A}}$ to configurations with counter value at most $K$, and by substituting all transitions leaving each $p(K)$ with a self-loop of the form $p(K) \rightsquigarrow p(K)$.

First, let us assume that $\bar{x} \geq 0$. By Proposition 5 (A), we have that $\mathrm{Val}(q(i)) = \infty$ for all $q \in Q$ and $i \geq |Q|$. Hence, it remains to approximate the value and compute $\varepsilon$-optimal strategy for all configurations $q(i)$ where $i \leq |Q|$. Actually, we can even compute these values precisely and construct a strategy $\hat{\sigma}$ which is optimal in each such $q(i)$. This is achieved simply by considering the finite-state MDP $\mathcal{G}_{|Q|}$ and solving the objective of minimizing the expected number of transitions needed to reach a state of the form $p(0)$, which can be done by standard methods in time polynomial in $||\mathcal{A}||$.

If $\bar{x} < 0$, we argue as follows. The strategy $\sigma$ of Proposition 5 (B.2) is not necessarily $\varepsilon$-optimal in $q(i)$, so we cannot use it directly. To overcome this problem, consider an *optimal* strategy $\pi^*$ in $q(i)$, and let $x_\ell$ be the probability that a run initiated in $q(i)$ (under the strategy $\pi^*$) visits a configuration of the form $r(i + \ell)$. Obviously, $x_\ell \cdot \min_{r\in Q}\{\mathbb{E}^{\pi^*} r(i{+}\ell)\} \;\leq\; \mathbb{E}^{\sigma} q(i)$, because otherwise $\pi^*$ would not be optimal in $q(i)$. Using the lower/upper bounds for $\mathbb{E}^{\pi^*} r(i{+}\ell)$ and $\mathbb{E}^{\sigma} q(i)$ given in Proposition 5 (B), we obtain $x_\ell \leq (i + U)/(i + \ell - V)$. Then, we compute $k \in \mathbb{N}$ such that

$$x_k \cdot \left(\max_{r\in Q}\left\{(i + k + U)/|\bar{x}| - \mathbb{E}^{\pi^*} r(i{+}k)\right\}\right) \quad\leq\quad \varepsilon$$

A simple computation reveals that it suffices to choose any $k$ such that

$$k \quad\geq\quad \frac{(i + U)\cdot(U + V)}{\varepsilon \cdot |\bar{x}|} + V - i,$$

[3] Recall that rational numbers are represented as fractions of binary numbers.

so the value of $k$ is exponential in $\|\mathcal{A}\|$ and polynomial in $i$ and $1/\varepsilon$. Now, consider $\mathcal{G}_{i+k}$, and let $f$ be a reward function over the transitions of $\mathcal{G}_{i+k}$ such that the loops on configurations where the counter equals 0 or $i + k$ have zero reward, transitions leading to configurations of the form $r(i{+}k)$ have reward $(i + k + U)/|\bar{x}|$, and all of the remaining transitions have reward 1. Now we solve the finite-state MDP $\mathcal{G}_{i+k}$ with the objective of minimizing the total accumulated reward. Note that an optimal strategy $\varrho$ in $\mathcal{G}_{i+k}$ is computable in time polynomial in the size of $\mathcal{G}_{i+k}$ [19]. Then, we define the corresponding strategy $\hat{\sigma}$ in $\mathcal{M}^{\infty}_{\mathcal{A}}$, which behaves like $\varrho$ until the counter reaches $i + k$, and from that point on it behaves like the counterless strategy $\sigma$. It is easy to see that $\hat{\sigma}$ is indeed $\varepsilon$-optimal in $q(i)$.

**Proof of Proposition 5.** Similarly as in [3], we use the solution $(\bar{x}, (\bar{z}_q)_{q\in Q}) \in \mathbb{Q}^{|Q|+1}$ of $\mathcal{L}$ to define a suitable submartingale, which is then used to derive the required bounds. In [3], Azuma's inequality was applied to the submartingale to prove exponential tail bounds for termination probability. In this paper, we need to use the optional stopping theorem rather than Azuma's inequality, and therefore we need to define the submartingale relative to a suitable filtration so that we can introduce an appropriate stopping time (without the filtration, the stopping time would have to depend just on numerical values returned by the martingale, which does not suit our purposes).

Given the solution $(\bar{x}, (\bar{z}_q)_{q\in Q}) \in \mathbb{Q}^{|Q|+1}$ from Lemma 4, we define a sequence of random variables $\{m^{(i)}\}_{i\geq 0}$ by setting

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

Note that for every history $u$ of length $i$ and every $0 \leq j \leq i$, the random variable $m^{(j)}$ returns the same value for every $\omega \in Run(u)$. The same holds for variables $S^{(j)}$ and $C^{(j)}$. We will denote these common values $m^{(j)}(u)$, $S^{(j)}(u)$ and $C^{(j)}(u)$, respectively. Using the same arguments as in Lemma 3 of [3], one may show that for every history $u$ of length $i$ we have $\mathbb{E}(m^{(i+1)} \mid Run(u)) \geq m^{(i)}(u)$. This shows that $\{m^{(i)}\}_{i\geq 0}$ is a *submartingale relative to the filtration* $\{\mathcal{F}_i\}_{i\geq 0}$, where for each $i \geq 0$ the $\sigma$-algebra $\mathcal{F}_i$ is the $\sigma$-algebra generated by all $Run(u)$ where $len(u) = i$. Intuitively, this means that value $m^{(i)}(\omega)$ is uniquely determined by prefix of $\omega$ of length $i$ and that the process $\{m^{(i)}\}_{i\geq 0}$ has nonnegative average change. For relevant definitions of (sub)martingales see, e.g., [21]. Another important observation is that $|m^{(i+1)} - m^{(i)}| \leq 1 + \bar{x} + V$ for every $i \geq 0$, i.e., the differences of the submartingale are bounded.

**Lemma 6.** *Under an arbitrary strategy $\pi$ and with an arbitrary initial configuration $q(j)$ where $j \geq 0$, the process $\{m^{(i)}\}_{i\geq 0}$ is a submartingale (relative to the filtration $\{\mathcal{F}_i\}_{i\geq 0}$) with bounded differences.*

*Part (A) of Proposition 5.* This part can be proved by a routine application of the optional stopping theorem to the martingale $\{m^{(i)}\}_{i\geq 0}$. Let $\bar{z}_{\max} := \max_{q\in Q} \bar{z}_q$, and consider a configuration $p(\ell)$ where $\ell + \bar{z}_p > \bar{z}_{\max}$. Let $\sigma$ be a strategy which is optimal in every configuration. Assume, for the sake of contradiction, that $\mathrm{Val}(p(\ell)) < \infty$.

Let us fix $k \in \mathbb{N}$ such that $\ell + \bar{z}_p < \bar{z}_{\max} + k$ and define a stopping time $\tau$ which returns the first point in time in which either $m^{(\tau)} \geq \bar{z}_{\max} + k$, or $m^{(\tau)} \leq \bar{z}_{\max}$. To apply the optional stopping theorem, we need to show that the expectation of $\tau$ is finite.

We argue that every configuration $q(i)$ with $i \geq 1$ satisfies the following: under the optimal strategy $\sigma$, a configuration with counter height $i-1$ is reachable from $q(i)$ in at most $|Q|^2$ steps (i.e., with a bounded probability). To see this, realize that for every configuration $r(j)$ there is a successor, say $r'(j')$, such that $\mathrm{Val}(r(j)) > \mathrm{Val}(r'(j'))$. Now consider a run $w$ initiated in $q(i)$ obtained by subsequently choosing successors with smaller and smaller values. Note that whenever $w(j)$ and $w(j')$ with $j < j'$ have the same control state, the counter height of $w(j')$ must be strictly smaller than the one of $w(j)$ because otherwise the strategy $\sigma$ could be improved (it suffices to behave in $w(j)$ as in $w(j')$). It follows that there must be $k \leq |Q|^2$ such that the counter height of $w(k)$ is $i-1$. From this we obtain that the expected value of $\tau$ is finite because the probability of terminating from any configuration with bounded counter height is bounded from zero. Now we apply the optional stopping theorem and obtain $\mathbb{P}^{\sigma}_{p(\ell)}(m^{(\tau)} \geq \bar{z}_{\max}+k) \geq c/(k+d)$ for suitable constants $c, d > 0$. As $m^{(\tau)} \geq \bar{z}_{\max} + k$ implies $C^{(\tau)} \geq k$, we obtain that

$$\mathbb{P}^{\sigma}_{p(\ell)}(T \geq k) \quad \geq \quad \mathbb{P}^{\sigma}_{p(\ell)}(C^{(\tau)} \geq k) \quad \geq \quad \mathbb{P}^{\sigma}_{p(\ell)}(m^{(\tau)} \geq \bar{z}_{\max} + k) \quad \geq \quad \frac{c}{k+d}$$

and thus

$$\mathbb{E}^{\sigma} p(\ell) \quad = \quad \sum_{k=1}^{\infty} \mathbb{P}^{\sigma}_{p(\ell)}(T \geq k) \quad \geq \quad \sum_{k=1}^{\infty} \frac{c}{k+d} \quad = \quad \infty$$

which contradicts our assumption that $\sigma$ is optimal and $\mathrm{Val}(p(\ell)) < \infty$.

It remains to show that $\mathrm{Val}(p(\ell)) = \infty$ even for $\ell = |Q|$. This follows from the following simple observation:

**Lemma 7.** *For all $q \in Q$ and $i \geq |Q|$ we have that* $\mathrm{Val}(q(i)) < \infty$ *iff* $\mathrm{Val}(q(|Q|)) < \infty$.

The "only if" direction of Lemma 7 is trivial. For the other direction, let $\mathcal{B}_k$ denote the set of all $p \in Q$ such that $\mathrm{Val}(p(k)) < \infty$. Clearly, $\mathcal{B}_0 = Q$, $\mathcal{B}_k \subseteq \mathcal{B}_{k-1}$, and one can easily verify that $\mathcal{B}_k = \mathcal{B}_{k+1}$ implies $\mathcal{B}_k = \mathcal{B}_{k+\ell}$ for all $\ell \geq 0$. Hence, $\mathcal{B}_{|Q|} = \mathcal{B}_{|Q|+\ell}$ for all $\ell$. Note that Lemma 7 holds for general OC-MDPs (i.e., we do not need to assume that $\mathcal{M}_{\mathcal{A}}$ is strongly connected).

*Part (B1) of Proposition 5.* Let $\pi$ be a strategy and $q(i)$ a configuration where $i \geq 0$. If $\mathbb{E}^{\pi} q(i) = \infty$, we are done. Now assume $\mathbb{E}^{\pi} q(i) < \infty$. Observe that for every $k \geq 0$ and every run $\omega$, the membership of $\omega$ into $\{T \leq k\}$ depends only on the finite prefix of $\omega$ of length $k$. This means that $T$ is a stopping time relative to filtration $\{\mathcal{F}_n\}_{n\geq 0}$. Since $\mathbb{E}^{\pi} q(i) < \infty$ and the submartingale $\{m^{(n)}\}_{n\geq 0}$ has bounded differences, we can apply the optional stopping theorem and obtain $\mathbb{E}^{\pi}(m^{(0)}) \leq \mathbb{E}^{\pi}(m^{(T)})$. But $\mathbb{E}^{\pi}(m^{(0)}) = i + \bar{z}_q$ and $\mathbb{E}^{\pi}(m^{(T)}) = \mathbb{E}^{\pi} \bar{z}_{S^{(T)}} + \mathbb{E}^{\pi} q(i) \cdot |\bar{x}|$. Thus, we get $\mathbb{E}^{\pi} q(i) \geq (i + \bar{z}_q - \mathbb{E}^{\pi} \bar{z}_{S^{(T)}})/|\bar{x}| \geq (i - V)/|\bar{x}|$.

*Part (B2) of Proposition 5.* First we show how to construct the desired strategy $\sigma$. Recall again the linear program $\mathcal{L}$ of Figure 1. We have already shown that this program has an optimal solution $\left(\bar{x}, (\bar{z}_q)_{q\in Q}\right) \in \mathbb{Q}^{|Q|+1}$, and we assume that $\bar{x} < 0$. By the strong duality theorem, this means that the linear program dual to $\mathcal{L}$ also has a feasible solution $\left((\bar{y}_q)_{q\in Q_0}, (\bar{y}_{(q,i,q')})_{q\in Q_1,(q,i,q')\in\delta}\right)$. Let

$$D = \{q \in Q_0 \mid \bar{y}_q > 0\} \cup \{q \in Q_1 \mid \bar{y}_{(q,i,q')} > 0 \text{ for some } (q, i, q') \in \delta\}.$$

By Corollary 8.8.8 of [19], the solution $\left((\bar{y}_q)_{q\in Q_0}, (\bar{y}_{(q,i,q')})_{q\in Q_1,(q,i,q')\in\delta}\right)$ can be chosen so that for every $q \in Q_1$ there is at most one transition $(q, i, q')$ with $\bar{y}_{(q,i,q')} > 0$. Following the construction given in Section 8.8 of [19], we define a counterless deterministic strategy $\sigma$ such that

- in a state $q \in D \cap Q_1$, the strategy $\sigma$ selects the transition $(q, i, q')$ with $\bar{y}_{(q,i,q')} > 0$;
- in the states outside $D$, the strategy $\sigma$ behaves like an optimal strategy for the objective of reaching the set $D$.

Clearly, the strategy $\sigma$ is computable in time polynomial in $\|\mathcal{A}\|$. In the full version of this paper [**?**], we show that $\sigma$ indeed satisfies Part (B.2) of Proposition 5.

### 3.2 General OC-MDP

In this section we prove Theorem 3 for general OC-MDPs, i.e., we drop the assumption that $\mathcal{M}_{\mathcal{A}}$ is strongly connected. We say that $C \subseteq Q$ is an *end component of* $\mathcal{A}$ if $C$ is strongly connected and for every $p \in C \cap Q_0$ we have that $\{q \in Q \mid p \rightsquigarrow q\} \subseteq C$. A *maximal end component (MEC) of* $\mathcal{A}$ is an end component of $\mathcal{A}$ which is maximal w.r.t. set inclusion. The set of all MECs of $\mathcal{A}$ is denoted by $\mathit{MEC}(\mathcal{A})$. Every $C \in \mathit{MEC}(\mathcal{A})$ determines a strongly connected OC-MDP $\mathcal{A}_C = (C, (C \cap Q_0, C \cap Q_1), \delta \cap (C \times \{+1, 0, -1\} \times C), \{P_q\}_{q\in C\cap Q_0})$. Hence, we may apply Proposition 5 to $\mathcal{A}_C$, and we use $\bar{x}_C$ and $V_C$ to denote the constants of Proposition 5 computed for $\mathcal{A}_C$.

*Part 1. of Theorem 3.* We show how to compute, in time polynomial in $\|\mathcal{A}\|$, the set $Q_{\mathit{fin}} = \{p \in Q \mid \mathrm{Val}(p(k)) < \infty \text{ for all } k \geq 0\}$. From this we easily obtain Part 1. of Theorem 3, because for every configuration $q(i)$ where $i \geq 0$ we have the following:

- if $i \geq |Q|$, then $\mathrm{Val}(q(i)) < \infty$ iff $q \in Q_{\mathit{fin}}$ (see Lemma 7);
- if $i < |Q|$, then $\mathrm{Val}(q(i)) < \infty$ iff the set $\{p(0) \mid p \in Q\} \cup \{p(|Q|) \mid p \in Q_{\mathit{fin}}\}$ can be reached from $q(i)$ with probability 1 in the finite-state MDP $\mathcal{G}_{|Q|}$ defined in Section 3.1 (here we again use Lemma 7).

So, it suffices to show how to compute the set $Q_{\mathit{fin}}$ in polynomial time.

**Proposition 8.** *Let $Q_{<0}$ be the set of all states from which the set $H = \{q \in Q \mid q$ belongs to a MEC $C$ satisfying $\bar{x}_C < 0\}$ is reachable with probability 1. Then $Q_{\mathit{fin}} = Q_{<0}$. Moreover, the membership to $Q_{<0}$ is decidable in time polynomial in $\|\mathcal{A}\|$.*

*Part 2. of Theorem 3.* First, we generalize Part (B) of Proposition 5 into the following:

**Proposition 9.** *For every $q \in Q_{\mathit{fin}}$ there is a number $t_q$ computable in time polynomial in $\|\mathcal{A}\|$ such that $-1 \leq t_q < 0$, $1/|t_q| \in \exp\left(\|\mathcal{A}\|^{O(1)}\right)$, and the following holds:*

*(A) There is a counterless strategy $\sigma$ and a number $U \in \exp(\|\mathcal{A}\|^{O(1)})$ such that for every configuration $q(i)$ where $q \in Q_{\mathit{fin}}$ and $i \geq 0$ we have that $\mathbb{E}^{\sigma} q(i) \leq i/|t_q| + U$. Moreover, both $\sigma$ and $U$ are computable in time polynomial in $\|\mathcal{A}\|$.*

*(B) There is a number $L \in \exp(||\mathcal{A}||^{O(1)})$ such that for every strategy $\pi$ and every configuration $q(i)$ where $i \geq |Q|$ we have that $\mathbb{E}^{\pi} \geq i/|t_q| - L$. Moreover, $L$ is computable in time polynomial in $||\mathcal{A}||$.*

Once the Proposition 9 is proved, we can compute an $\varepsilon$-optimal strategy for an arbitrary configuration $q(i)$ where $q \in Q_{\mathit{fin}}$ and $i \geq |Q|$ in exactly the same way (and with the same complexity) as in the strongly connected case. Actually, it can also be used to compute the approximate values and $\varepsilon$-optimal strategies for configurations $q(j)$ such that $q \notin Q_{\mathit{fin}}$ or $1 \leq j < |Q|$. Observe that

- if $q \notin Q_{\mathit{fin}}$ and $j \geq |Q|$, the value is infinite by Part 1;
- otherwise, we construct the finite-state MDP $\mathcal{G}_{|Q|}$ (see Section 3.1) where the loops on configurations with counter value 0 have reward 0, the loops on configurations of the form $r(|Q|)$ have reward 0 or 1, depending on whether $r \in Q_{fin}$ or not, transitions leading to $r(|Q|)$ where $r \in Q_{fin}$ are rewarded with some $\varepsilon$-approximation of $\mathrm{Val}(r(|Q|))$, and all other transitions have reward 1. The reward function can be computed in time exponential in $||\mathcal{A}||$ by Proposition 9, and the minimal total accumulated reward from $q(j)$ in $\mathcal{G}_{|Q|}$, which can be computed by standard algorithms, is an $\varepsilon$-approximation of $\mathrm{Val}(q(j))$. The corresponding $\varepsilon$-optimal strategy can be computed in the obvious way.

The missing proofs of Propositions 8 and 9 can be found in [**?**].

## 4 Lower Bounds

In this section, we show that approximating $\mathrm{Val}(q(i))$ is computationally hard, even if $i = 1$ and the edge probabilities in the underlying OC-MDP are all equal to $1/2$. More precisely, we prove the following:

**Theorem 10.** *The value of a given configuration $q(1)$ cannot be approximated in polynomial time up to a given absolute/relative error $\varepsilon > 0$ unless P=NP, even if all outgoing edges of all stochastic control states in the underlying OC-MDP have probability $1/2$.*

The proof of Theorem 10 is split into two phases, which are relatively independent. First, we show that given a propositional formula $\varphi$, one can efficiently compute an OC-MDP $\mathcal{A}$, a configuration $p(K)$ of $\mathcal{A}$, and a number $N$ such that the value of $p(K)$ is either $N - 1$ or $N$ depending on whether $\varphi$ is satisfiable or not, respectively. The numbers $K$ and $N$ are exponential in $||\varphi||$, which means that their encoding size is polynomial (we represent all numerical constants in binary). Here we use the technique of encoding propositional assignments into counter values presented in [17], but we also need to invent some specific gadgets to deal with our specific objective. The first part already implies that approximating $\mathrm{Val}(q(i))$ is computationally hard. In the second phase, we show that the same holds also for configurations where the counter is initiated to 1. This is achieved by employing another gadget which just increases the counter to an exponentially high value with a sufficiently large probability. The two phases are elaborated in [**?**].

## References

1. Proceedings of FST&TCS 2010, LIPIcs, vol. 8. Schloss Dagstuhl (2010)
2. Brázdil, T., Brožek, V., Etessami, K.: One-counter stochastic games. In: Proceedings of FST&TCS 2010 [1], pp. 108–119
3. Brázdil, T., Brožek, V., Etessami, K., Kučera, A.: Approximating the termination value of one-counter MDPs and stochastic games. In: Aceto, L., Henzinger, M., Sgall, J. (eds.) Proceedings of ICALP 2011, Part II. LNCS, vol. 6756, pp. 332–343. Springer (2011)
4. Brázdil, T., Brožek, V., Etessami, K., Kučera, A., Wojtczak, D.: One-counter Markov decision processes. In: Proceedings of SODA 2010. pp. 863–874. SIAM (2010)
5. Brázdil, T., Brožek, V., Forejt, V., Kučera, A.: Reachability in recursive Markov decision processes. I&C 206(5), 520–537 (2008)
6. Brázdil, T., Brožek, V., Kučera, A., Obdržálek, J.: Qualitative reachability in stochastic BPA games. I&C 208(7), 772–796 (2010)
7. Chatterjee, K., Doyen, L.: Energy parity games. In: Abramsky, S., Gavoille, C., Kirchner, C., Meyer auf der Heide, F., Spirakis, P. (eds.) Proceedings of ICALP 2010, Part II. LNCS, vol. 6199, pp. 599–610. Springer (2010)
8. Chatterjee, K., Doyen, L., Henzinger, T., Raskin, J.F.: Generalized mean-payoff and energy games. In: Proceedings of FST&TCS 2010 [1], pp. 505–516
9. Etessami, K., Wojtczak, D., Yannakakis, M.: Recursive stochastic games with positive rewards. In: Aceto, L., Damgard, I., Goldberg, L., Haldórsson, M., Ingólfsdóttir, A., Walukiewicz, I. (eds.) Proceedings of ICALP 2008, Part I. LNCS, vol. 5125, pp. 711–723. Springer (2008)
10. Etessami, K., Wojtczak, D., Yannakakis, M.: Quasi-birth-death processes, tree-like QBDs, probabilistic 1-counter automata, and pushdown systems. Performance Evaluation 67(9), 837–857 (2010)
11. Etessami, K., Yannakakis, M.: Recursive Markov decision processes and recursive stochastic games. In: Caires, L., Italiano, G., Monteiro, L., Palamidessi, C., Yung, M. (eds.) Proceedings of ICALP 2005. LNCS, vol. 3580, pp. 891–903. Springer (2005)
12. Etessami, K., Yannakakis, M.: Efficient qualitative analysis of classes of recursive Markov decision processes and simple stochastic games. In: Durand, B., Thomas, W. (eds.) Proceedings of STACS 2006. LNCS, vol. 3884, pp. 634–645. Springer (2006)
13. Filar, J., Vrieze, K.: Competitive Markov Decision Processes. Springer (1996)
14. Göller, S., Lohrey, M.: Branching-time model checking of one-counter processes. In: Proceedings of STACS 2010. LIPIcs, vol. 5, pp. 405–416. Schloss Dagstuhl (2010)
15. Jančar, P., Kučera, A., Moller, F., Sawa, Z.: DP lower bounds for equivalence-checking and model-checking of one-counter automata. I&C 188(1), 1–19 (2004)
16. Jančar, P., Sawa, Z.: A note on emptiness for alternating finite automata with a one-letter alphabet. IPL 104(5), 164–167 (2007)
17. Kučera, A.: The complexity of bisimilarity-checking for one-counter processes. TCS 304(1–3), 157–183 (2003)
18. Latouche, G., Ramaswami, V.: Introduction to Matrix Analytic Methods in Stochastic Modeling. ASA-SIAM series on statistics and applied probability (1999)
19. Puterman, M.: Markov Decision Processes. Wiley (1994)
20. Serre, O.: Parity games played on transition graphs of one-counter processes. In: Aceto, L., Ingólfsdóttir, A. (eds.) Proceedings of FoSSaCS 2006. LNCS, vol. 3921, pp. 337–351. Springer (2006)
21. Williams, D.: Probability with Martingales. Cambridge University Press (1991)