Strategy Complexity of Mean Payoff, Total Payoff and Point Payoff Objectives in Countable MDPs

Richard Mayr
University of Edinburgh, UK

Eric Munday
University of Edinburgh, UK

Abstract

We study countably infinite Markov decision processes (MDPs) with real-valued transition rewards. Every infinite run induces the following sequences of payoffs: 1. Point payoff (the sequence of directly seen transition rewards), 2. Total payoff (the sequence of the sums of all rewards so far), and 3. Mean payoff. For each payoff type, the objective is to maximize the probability that the lim inf is non-negative. We establish the complete picture of the strategy complexity of these objectives, i.e., how much memory is necessary and sufficient for $\varepsilon$-optimal (resp. optimal) strategies. Some cases can be won with memoryless deterministic strategies, while others require a step counter, a reward counter, or both.

2012 ACM Subject Classification Theory of computation → Random walks and Markov chains; Mathematics of computing → Probability and statistics

Keywords and phrases Markov decision processes, Strategy complexity, Mean payoff


Acknowledgements We thank an anonymous reviewer for very detailed and helpful comments.

1 Introduction

Background. Markov decision processes (MDPs) are a standard model for dynamic systems that exhibit both stochastic and controlled behavior [18]. Applications include control theory [5, 1], operations research and finance [2, 6, 20], artificial intelligence and machine learning [23, 21], and formal verification [9, 3].

An MDP is a directed graph where states are either random or controlled. In a random state the next state is chosen according to a fixed probability distribution. In a controlled state the controller can choose a distribution over all possible successor states. By fixing a strategy for the controller (and an initial state), one obtains a probability space of runs of the MDP. The goal of the controller is to optimize the expected value of some objective function on the runs. The type of strategy necessary to achieve an $\varepsilon$-optimal (resp. optimal) value for a given objective is called its strategy complexity.

Transition rewards and liminf objectives. MDPs are given a reward structure by assigning a real-valued (resp. integer or rational) reward to each transition. Every run then induces an infinite sequence of seen transition rewards $r_0r_1r_2\ldots$. We consider the lim inf of this sequence, as well as two other important derived sequences.

1. The point payoff considers the lim inf of the sequence $r_0r_1r_2\ldots$ directly.
2. The total payoff considers the lim inf of the sequence $\left\{ \sum_{i=0}^{n-1} r_i \right\}_{n \in \mathbb{N}}$, i.e., the sum of all rewards seen so far.
3. The mean payoff considers the lim inf of the sequence $\left\{ \frac{1}{n} \sum_{i=0}^{n-1} r_i \right\}_{n \in \mathbb{N}}$, i.e., the mean of all rewards seen so far in an expanding prefix of the run.

For each of the three cases above, the lim inf threshold objective is to maximize the probability that the lim inf of the respective type of sequence is $\geq 0$. 

© Richard Mayr and Eric Munday; licensed under Creative Commons License CC-BY 4.0
32nd International Conference on Concurrency Theory (CONCUR 2021).
Editors: Serge Haddad and Daniele Varacca; Article No. 12; pp. 12:1–12:15
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany
Our contribution. We establish the strategy complexity of all the lim inf threshold objectives above for countably infinite MDPs. (For the simpler case of finite MDPs, see the paragraph on related work below.) We show the amount and type of memory that is sufficient for ε-optimal strategies (and optimal strategies, where they exist), and corresponding lower bounds in the sense of Remark 1. This is not only the distinction between memoryless, finite memory and infinite memory, but the type of infinite memory that is necessary and sufficient. A step counter is an integer counter that merely counts the number of steps in the run (i.e., like a discrete clock), while a reward counter is a variable that records the sum of all rewards seen so far. (The reward counter has the same type as the transition rewards in the MDP, i.e., integers, rationals or reals.) While these use infinite memory, it is a very restricted form, since this memory is not directly controlled by the player. Strategies using only a step counter are also called Markov strategies [18].

Some of the lim inf objectives can be attained by memoryless deterministic (MD) strategies, while others require (in the sense of Remark 1) a step counter, a reward counter, or both. It depends on the type of objective (point, total, or mean payoff) and on whether the MDP is finitely or infinitely branching. For clarity of presentation, our counterexamples use large transition rewards and high degrees of branching. However, the lower bounds hold even for just binary branching MDPs with transition rewards in \{-1, 0, 1\}; cf. [17].

For our objectives, the strategy complexities of ε-optimal and optimal strategies (where they exist) coincide, but the proofs are different. Table 1 shows the results for all combinations.

<table>
<thead>
<tr>
<th></th>
<th>Point payoff</th>
<th>Total payoff</th>
<th>Mean payoff</th>
</tr>
</thead>
<tbody>
<tr>
<td>ε-optimal, infinitely branching</td>
<td>SC 17, 32</td>
<td>SC+RC 17, 9, 34</td>
<td>SC+RC 15, 8, 33</td>
</tr>
<tr>
<td>optimal, infinitely branching</td>
<td>SC 17, 35</td>
<td>SC+RC 14, 17, 35</td>
<td>SC+RC 13, 16, 35</td>
</tr>
<tr>
<td>ε-optimal, finitely branching</td>
<td>MD 27</td>
<td>RC 9, 30</td>
<td>SC+RC 15, 8, 33</td>
</tr>
<tr>
<td>optimal, finitely branching</td>
<td>MD 31</td>
<td>RC 14, 31</td>
<td>SC+RC 13, 16, 35</td>
</tr>
</tbody>
</table>

Some complex new proof techniques are developed to show these results. E.g., the examples showing the lower bound in cases where both a step counter and a reward counter are required use a finely tuned tradeoff between different risks that can be managed with both counters, but not with just one counter plus arbitrary finite memory. The strategies showing the upper bounds need to take into account convergence effects, e.g., the sequence of point rewards \(-1/2, -1/3, -1/4, \ldots\) does satisfy \(\lim \inf \geq 0\), i.e., one cannot assume that rewards are integers.

Due to space constraints, we sketch some proofs in the main body. Full proofs can be found in [17].

Related work. Mean payoff objectives for finite MDPs have been widely studied; cf. survey in [8]. There exist optimal MD strategies for lim inf mean payoff (which are also optimal for lim sup mean payoff since the transition rewards are bounded), and the associated computational problems can be solved in polynomial time [8, 18]. Similarly, see [7] for a survey on lim sup and lim inf point payoff objectives in finite stochastic games and MDPs, where there also exist optimal MD strategies, and the more recent paper by Flesch, Predtetchinski and Sudderth [11] on simplifying optimal strategies.
All this does not carry over to countably infinite MDPs. Optimal strategies need not exist (not even for much simpler objectives). ($\varepsilon$-)optimal strategies can require infinite memory, and computational problems are not defined in general, since a countable MDP need not be finitely presented [16]. Moreover, attainment for lim inf mean payoff need not coincide with attainment for lim sup mean payoff, even for very simple examples. E.g., consider the acyclic infinite graph with transitions $s_n \rightarrow s_{n+1}$ for all $n \in \mathbb{N}$ with reward $(-1)^n2^n$ in the $n$-th step, which yields a lim inf mean payoff of $-\infty$ and a lim sup mean payoff of $+\infty$.

Mean payoff objectives for countably infinite MDPs have been considered in [18, Section 8.10], e.g., [18, Example 8.10.2] shows that there are no optimal MDP (memoryless deterministic) strategies for lim inf/lim sup mean payoff. [19, Counterexample 1.3] shows that there are not even $\varepsilon$-optimal memoryless randomized strategies for lim inf/lim sup mean payoff. (We show much stronger lower/upper bounds; cf. Table 1.)

Sudderth [22] considered an objective on countable MDPs that is related to our point payoff threshold objective. However, instead of maximizing the probability that the lim inf/lim sup is non-negative, it asks to maximize the threshold objective. However, instead of maximizing the probability that the remaining cases have a huge positive mean payoff, even for very simple examples. E.g., consider the mean payoff objectives for countably infinite MDPs have been considered in [18, Section 8.10], e.g., [18, Example 8.10.2] shows that there are no optimal MDP (memoryless deterministic) strategies for lim inf/lim sup mean payoff. [19, Counterexample 1.3] shows that there are not even $\varepsilon$-optimal memoryless randomized strategies for lim inf/lim sup mean payoff. (We show much stronger lower/upper bounds; cf. Table 1.)

Sudderth [22] considered an objective on countable MDPs that is related to our point payoff threshold objective. However, instead of maximizing the probability that the lim inf/lim sup is non-negative, it asks to maximize the expectation of the lim inf/lim sup point payoffs, which is a different problem (e.g., it can tolerate a high probability of a negative lim inf/lim sup if the remaining cases have a huge positive lim inf/lim sup). Hill & Pestien [12] showed the existence of good randomized Markov strategies for the lim sup of the expected average reward up-to step $n$ for growing $n$, and for the expected lim inf of the point payoffs.

## 2 Preliminaries

Markov decision processes. A probability distribution over a countable set $S$ is a function $f : S \rightarrow [0, 1]$ with $\sum_{s \in S} f(s) = 1$. We write $\mathcal{D}(S)$ for the set of all probability distributions over $S$. A Markov decision process (MDP) $\mathcal{M} = (S, S_0, S_\infty, \rightarrow, P, r)$ consists of a countable set $S$ of states, which is partitioned into a set $S_0$ of controlled states and a set $S_\infty$ of random states, a transition relation $\rightarrow \subseteq S \times S$, and a probability function $P : S_\infty \rightarrow \mathcal{D}(S)$. We write $s \rightarrow s'$ if $(s, s') \in \rightarrow$, and refer to $s'$ as a successor of $s$. We assume that every state has at least one successor. The probability function $P$ assigns to each random state $s \in S_\infty$ a probability distribution $P(s)$ over its (non-empty) set of successor states. A sink in $\mathcal{M}$ is a subset $T \subseteq S$ closed under the $\rightarrow$ relation, that is, $s \in T$ and $s \rightarrow s'$ implies that $s' \in T$.

An MDP is acyclic if the underlying directed graph $(S, \rightarrow)$ is acyclic, i.e., there is no directed cycle. It is finitely branching if every state has finitely many successors and infinitely branching otherwise. An MDP without controlled states ($S_0 = \emptyset$) is called a Markov chain.

In order to specify our mean/total/point payoff objectives (see below), we define a function $r : S \times S \rightarrow \mathbb{R}$ that assigns numeric rewards to transitions.

Strategies and Probability Measures. A run $\rho$ is an infinite sequence of states and transitions $s_0s_1e_1 \cdots$ such that $e_i = (s_i, s_{i+1}) \in \rightarrow$ for all $i \in \mathbb{N}$. Let $\text{Runs}_\mathcal{M}$ be the set of all runs from $s_0$ in the MDP $\mathcal{M}$. A partial run is a finite prefix of a run, $p\text{Runs}_\mathcal{M}$ is the set of all partial runs from $s_0$ and $p\text{Runs}_\mathcal{M}$ the set of partial runs from any state.

We write $\rho(i) \overset{def}{=} s_i$ for the $i$-th state along $\rho$ and $\rho(i) \overset{def}{=} e_i$ for the $i$-th transition along $\rho$. We sometimes write runs as $s_0s_1 \cdots$, leaving the transitions implicit. We say that a (partial) run $\rho$ visits $s$ if $s = \rho(i)$ for some $i$, and that $\rho$ starts in $s$ if $s = \rho(0)$.

A strategy is a function $\sigma : \text{pRuns}_\mathcal{M} : S_0 \rightarrow \mathcal{D}(S)$ that assigns to partial runs $\rho$s, where $s \in S_0$, a distribution over the successors $\{s' \in S \mid s \rightarrow s'\}$. The set of all strategies in $\mathcal{M}$ is denoted by $\Sigma_\mathcal{M}$ (we omit the subscript and write $\Sigma$ if $\mathcal{M}$ is clear from the context). A (partial) run $s_0s_1e_1 \cdots$ is consistent with a strategy $\sigma$ if for all $i$ either $s_i \in S_0$ and $\sigma(s_0s_1e_1 \cdots s_i)(s_{i+1}) > 0$, or $s_i \in S_\infty$ and $P(s_i)(s_{i+1}) > 0$. 

CONCUR 2021
An MDP $M = (S, S_0, S_f, \rightarrow, P, r)$, an initial state $s_0 \in S$, and a strategy $\sigma$ induce a probability space in which the outcomes are runs starting in $s_0$ and with measure $P_{M,s_0,\sigma}$ defined as follows. It is first defined on cylinders $s_0 \sigma_0 s_1 \sigma_1 \ldots$, where $s_0 \sigma_0 s_1 \sigma_1 \ldots$ is a run starting in $s_0$. If $s_0 \sigma_0 s_1 \sigma_1 \ldots$ is not a partial run consistent with $\sigma$, then $P_{M,s_0,\sigma}(s_0 \sigma_0 s_1 \sigma_1 \ldots ) = 0$. Otherwise, $P_{M,s_0,\sigma}(s_0 \sigma_0 s_1 \sigma_1 \ldots ) = \prod_{i=0}^{n-1} \delta(s_0 \sigma_0 s_1 \ldots s_i)(s_{i+1})$, where $\delta$ is the map that extends $\sigma$ by $\delta(ws) = P(s)$ for all partial runs $ws \in p Runs_M$. By Carathéodory’s theorem [4], this extends uniquely to a probability measure $P_{M,s_0,\sigma}$ on the Borel $\sigma$-algebra $F$ of subsets of $Run_M$. Elements of $F$, i.e., measurable sets of runs, are called events or objectives here. For $X \in F$ we will write $X \equiv Run_M^1 \setminus X \in F$ for its complement and $E_{M,s_0,\sigma}$ for the expectation w.r.t. $P_{M,s_0,\sigma}$. We drop the indices if possible without ambiguity.

Objectives. We consider objectives that are determined by a predicate on infinite runs. We assume familiarity with the syntax and semantics of the temporal logic LTL [10]. Formulas are interpreted on the structure $(S, \rightarrow)$. We use $[\varphi]^s$ to denote the set of runs starting from $s$ that satisfy the LTL formula $\varphi$, which is a measurable set [24]. We also write $[[\varphi]]$ for $\bigcup_{s \in S} [\varphi]^s$. Where it does not cause confusion we will identify $\varphi$ and $[\varphi]$ and just write $P_{M,s,\sigma}(\varphi)$ instead of $P_{M,s,\sigma}([\varphi]^s)$. The reachability objective of eventually visiting a set of states $X$ can be expressed by $[FX] \equiv \{ \rho | \exists i, \rho_\pi(i) \in X \}$. Reaching $X$ within at most $k$ steps is expressed by $[F^k X] \equiv \{ \rho | \exists i \leq k, \rho_\pi(i) \in X \}$. The definitions for eventually visiting certain transitions are analogous. The operator $G$ (always) is defined as $-F\neg$. So the safety objective of avoiding $X$ is expressed by $G\neg X$.

- The $PP_{liminf \geq 0}$ objective is to maximize the probability that the lim inf of the point payoffs (the immediate transition rewards) is $\geq 0$, i.e., $PP_{liminf \geq 0} \equiv \{ \rho | \liminf_{n \in \mathbb{N}} r(\rho_\pi(n)) \geq 0 \}$.
- The $TP_{liminf \geq 0}$ objective is to maximize the probability that the lim inf of the total payoff (the sum of the transition rewards seen so far) is $\geq 0$, i.e., $TP_{liminf \geq 0} \equiv \{ \rho | \liminf_{n \in \mathbb{N}} \sum_{j=0}^{n-1} r(\rho_\pi(j)) \geq 0 \}$.
- The $MP_{liminf \geq 0}$ objective is to maximize the probability that the lim inf of the mean payoff is $\geq 0$, i.e., $MP_{liminf \geq 0} \equiv \{ \rho | \liminf_{n \in \mathbb{N}} \frac{1}{n} \sum_{j=0}^{n-1} r(\rho_\pi(j)) \geq 0 \}$.

An objective $\varphi$ is called tail in $M$ if for every run $\rho' \rho$ in $M$ with some finite prefix $\rho'$ we have $\rho' \rho \in [\varphi] \Leftrightarrow \rho \in [\varphi]$. An objective is called a tail objective if it is tail in every MDP. $PP_{liminf \geq 0}$ and $MP_{liminf \geq 0}$ are tail objectives, but $TP_{liminf \geq 0}$ is not. Also $PP_{liminf \geq 0}$ is more general than co-Büchi. (The special case of integer transition rewards coincides with co-Büchi, since rewards $\leq -1$ and accepting states can be encoded into each other.)

Strategy Classes. Strategies are in general randomized (R) in the sense that they take values in $D(S)$. A strategy $\sigma$ is deterministic (D) if $\sigma(\rho)$ is a Dirac distribution for all $\rho$. General strategies can be history dependent (H), while others are restricted by the size or type of memory they use, see below. We consider certain classes of strategies:

- A strategy $\sigma$ is memoryless (M) (also called positional) if it can be implemented with a memory of size 1. We may view M-strategies as functions $\sigma : S_0 \rightarrow D(S)$.
- A strategy $\sigma$ is finite memory (F) if there exists a finite memory $M$ implementing $\sigma$.
- Hence FR stands for finite memory randomized.
- A step counter strategy bases decisions only on the current state and the number of steps taken so far, i.e., it uses an unbounded integer counter that gets incremented by 1 in every step. Such strategies are also called Markov strategies [18].
- $k$-bit Markov strategies use $k$ extra bits of general purpose memory in addition to a step counter [15].
- A **reward counter strategy** uses infinite memory, but only in the form of a counter that always contains the sum of all transition rewards seen so far.
- A **step counter + reward counter strategy** uses both a step counter and a reward counter.

See [17] for a formal definition how strategies use memory. Step counters and reward counters are very restricted forms of memory, since the memory update is not directly under the control of the player. These counters merely record an aspect of the partial run.

Optimal and $\epsilon$-optimal Strategies. Given an objective $\varphi$, the value of state $s$ in an MDP $M$, denoted by $\text{val}_{M,\varphi}(s)$, is the supremum probability of achieving $\varphi$. Formally, $\text{val}_{M,\varphi}(s) \overset{\text{def}}{=} \sup_{\sigma \in \Sigma} P_{M,s,\sigma}(\varphi)$ where $\Sigma$ is the set of all strategies. For $\epsilon \geq 0$ and state $s \in S$, we say that a strategy is $\epsilon$-optimal from $s$ if $P_{M,s,\sigma}(\varphi) \geq \text{val}_{M,\varphi}(s) - \epsilon$. A 0-optimal strategy is called **optimal**. An optimal strategy is *almost-surely winning* if $\text{val}_{M,\varphi}(s) = 1$. Considering an MD strategy as a function $\sigma : S_0 \rightarrow S$ and $\epsilon \geq 0$, $\sigma$ is uniformly $\epsilon$-optimal (resp. uniformly optimal) if it is $\epsilon$-optimal (resp. optimal) from every $s \in S$.

**Remark 1.** To establish an upper bound $X$ on the strategy complexity of an objective $\varphi$ in countable MDPs, it suffices to prove that there always exist good ($\epsilon$-optimal, resp. optimal) strategies in class $X$ (e.g., MD, MR, FD, FR, etc.) for objective $\varphi$.

Lower bounds on the strategy complexity of an objective $\varphi$ can only be established in the sense of proving that good strategies for $\varphi$ do not exist in some classes $Y$, $Z$, etc. Classes of strategies that use different types of restricted infinite memory are generally not comparable, e.g., step counter strategies are incomparable to reward counter strategies. In particular, there is no weakest type of infinite memory with restricted use. Therefore statements like “good strategies for objective $\varphi$ require at least a step counter” are always relative to the considered alternative strategy classes. In this paper, we only consider the strategy classes of memoryless, finite memory, step counter, reward counter and combinations thereof. Thus, when we write in Table 1 that an objective requires a step counter (SC), it just means that a reward counter (RC) plus finite memory is not sufficient.

For our upper bounds, we use deterministic strategies. Moreover, we show that allowing randomization does not help to reduce the strategy complexity, in the sense of Remark 1.

### 3 When is a step counter not sufficient?

In this section we will prove that strategies with a step counter plus arbitrary finite memory are not sufficient for $\epsilon$-optimal strategies for $MP_{\text{lim inf} \geq 0}$ or $TP_{\text{lim inf} \geq 0}$. We will construct an acyclic MDP where the step counter is implicit in the state such that $\epsilon$-optimal strategies for $MP_{\text{lim inf} \geq 0}$ and $TP_{\text{lim inf} \geq 0}$ still require infinite memory.

#### 3.1 Epsilon-optimal strategies

We construct an acyclic MDP $M$ in which the step counter is implicit in the state as follows.

The system consists of a sequence of gadgets. Figure 1 depicts a typical building block in this system. The system consists of these gadgets chained together as illustrated in Figure 2, starting with $n$ sufficiently high at $n = N^\ast$. In the controlled choice, there is a small chance in all but the top choice of falling into a $\perp$ state. These $\perp$ states are abbreviations for an infinite chain of states with $-1$ reward on the transitions and are thus losing. The intuition behind the construction is that there is a random transition with branching degree $k(n) + 1$. Then, the only way to win, in the controlled states, is to play the $i$-th choice if one arrived from the $i$-th choice. Thus intuitively, to remember what this choice was, one requires at least $k(n) + 1$ memory modes. That is to say, the one and only way to win is to mimic, and mimicry requires memory.
Additionally, the related family of series $\delta_k(n)$ and $\varepsilon_i(n)$ are probabilities depending on $n$ and the $\pm m_n$ are transition rewards. We index the successor states of $s_n$ and $e_n$ from $0$ to $k(n)$ to match the indexing of the $\delta$'s and $\varepsilon$'s such that the bottom state is indexed with $0$ and the top state with $k(n)$.

**Remark 2.** $M$ is acyclic, finitely branching and for every state $s \in S$, $\exists n_s \in \mathbb{N}$ such that every path from $s_0$ to $s$ has length $n_s$. That is to say the step counter is implicit in the state. Additionally, the number of transitions in each gadget now grows unboundedly with $n$ according to the function $k(n)$. Consequently, we will show that the number of memory modes required to play correctly grows above every finite bound. This will imply that no finite amount of memory suffices for $\varepsilon$-optimal strategies.

**Notation.** All logarithms are assumed to be in base $e$.

$$
\begin{align*}
\log_i n & \overset{\text{def}}{=} \log n, \quad \log_{i+1} n \overset{\text{def}}{=} \log \log_i n \\
\delta_0(n) & \overset{\text{def}}{=} \frac{1}{\log n}, \quad \delta_i(n) \overset{\text{def}}{=} \frac{1}{\log_{i+1} n}, \quad \delta_{k(n)}(n) \overset{\text{def}}{=} 1 - \sum_{j=0}^{k(n)-1} \delta_j(n) \\
\varepsilon_0(n) & \overset{\text{def}}{=} \frac{1}{\log n}, \quad \varepsilon_{i+1}(n) \overset{\text{def}}{=} \frac{\varepsilon_i(n)}{\log_{i+2} n}, \quad \text{i.e.} \quad \varepsilon_i(n) = \frac{1}{n \cdot \log n \cdot \log_2 n \cdot \ldots \cdot \log_{i+1} n}, \quad \varepsilon_{k(n)}(n) \overset{\text{def}}{=} 0 \\
\text{Tower}(0) & \overset{\text{def}}{=} e^0 = 1, \quad \text{Tower}(i+1) \overset{\text{def}}{=} e^{\text{Tower}(i)}, \quad N_i \overset{\text{def}}{=} \text{Tower}(i)
\end{align*}
$$

**Lemma 3.** The family of series $\sum_{i \geq N} \delta_j(n) \cdot \varepsilon_i(n)$ is divergent for all $i, j \in \mathbb{N}, i < j$. Additionally, the related family of series $\sum_{i \geq N_i} \delta_i(n) \cdot \varepsilon_i(n)$ is convergent for all $i \in \mathbb{N}$.

**Proof.** These are direct consequences of Cauchy’s Condensation Test.

**Definition 4.** We define $k(n)$, the rate at which the number of transitions grows. We define $k(n)$ in terms of fast growing functions $g$, Tower and $h$ defined for $i \geq 1$ as follows:

$$
g(i) \overset{\text{def}}{=} \min \left\{ N : \left( \sum_{n > N} \delta_{i-1}(n) \varepsilon_{i-1}(n) \right) \leq 2^{-i} \right\}, \quad h(1) \overset{\text{def}}{=} 2
$$
ensure that the transition probabilities are always well defined. That cannot fail unless it hits a ⊥ state. The proofs of e.g. Lemma 6.

\[ h(i + 1) \overset{\text{def}}{=} \max \left\{ g(i + 1), \text{Tower}(i + 2), \min \left\{ m + 1 \in \mathbb{N} : \sum_{n=h(i)}^{m} \varepsilon_{i-1}(n) \geq 1 \right\} \right\}. \]

Note that function \( g \) is well defined by Lemma 3, and \( h(i + 1) \) is well defined since for all \( i \), \( \sum_{n=h(i)}^{\infty} \varepsilon_{i-1}(n) \) diverges to infinity. \( k(n) \) is a slow growing unbounded step function defined in terms of \( h \) as \( k(n) \overset{\text{def}}{=} h^{-1}(n) \). The Tower function features in the definition to ensure that the transition probabilities are always well defined. \( g \) and \( h \) are used to smooth the proofs of e.g. Lemma 6. Notation: \( N^* \overset{\text{def}}{=} \min \{ n \in \mathbb{N} : k(n) = 1 \} \). This is intuitively the first natural number for which the construction is well defined.

The reward \( m_n \) which appears in the \( n \)-th gadget is defined such that it outweighs any possible reward accumulated up to that point in previous gadgets. As such we define \( m_n \overset{\text{def}}{=} 2k(n) \sum_{i=N^*}^{n-1} m_i \), with \( m_{N^*} \overset{\text{def}}{=} 1 \) and where \( k(n) \) is the branching degree.

To simplify the notation, the state \( s_0 \) in our theorem statements refers to \( s_{N^*} \).

\textbf{Lemma 5.} For \( k(n) \geq 1 \), the transition probabilities in the gadgets are well defined.

\textbf{Lemma 6.} For every \( \varepsilon > 0 \), there exists a strategy \( \sigma_\varepsilon \) with \( \mathcal{P} \cdot s_0, \sigma_\varepsilon(MP_{\text{lim inf} \geq 0}) \geq 1 - \varepsilon \) that cannot fail unless it hits a ⊥ state. Formally, \( \mathcal{P} \cdot s_0, \sigma_\varepsilon(MP_{\text{lim inf} \geq 0} \land G(\neg \bot)) = \mathcal{P} \cdot s_0, \sigma_\varepsilon(G(\neg \bot)) \geq 1 - \varepsilon \). So in particular, \( \text{val}_{M, MP_{\text{lim inf} \geq 0}}(s_0) = 1 \).

\textbf{Proof sketch.} We define a strategy \( \sigma \) which in \( c_n \) always mimics the choice in \( s_n \). Playing according to \( \sigma \), the only way to lose is by dropping into the ⊥ state. This is because by mimicking, the player finishes each gadget with a reward of 0. From \( s_0 \), the probability of surviving while playing in all the gadgets is

\[ \prod_{n \geq N^*} \left( 1 - \sum_{j=0}^{k(n)-1} \delta_j(n) \cdot \varepsilon_j(n) \right) > 0. \]

Hence the player has a non zero chance of winning when playing \( \sigma \).

When playing with the ability to skip gadgets, as illustrated in Figure 2, all runs not visiting a ⊥ state are winning since the total reward never dips below 0. We then consider the strategy \( \sigma_\varepsilon \) which plays like \( \sigma \) after skipping forwards by sufficiently many gadgets (starting at \( n \geq N^* \)). Its probability of satisfying \( MP_{\text{lim inf} \geq 0} \) corresponds to a tail of the above product, which can be made arbitrarily close to 1 (and thus \( \geq 1 - \varepsilon \)). Thus the strategies \( \sigma_\varepsilon \) for arbitrarily small \( \varepsilon > 0 \) witness that \( \text{val}_{M, MP_{\text{lim inf} \geq 0}}(s_0) = 1 \).
\textbf{Lemma 7.} For any FR strategy \(\sigma\), almost surely either the mean payoff dips below \(-1\) infinitely often, or the run hits a \(\perp\) state, i.e. \(\mathcal{P}_{M,\sigma,s_0}(\text{MP}_{\lim \inf \geq 0}) = 0\).

\textbf{Proof sketch.} Let \(\sigma\) be some FR strategy with \(k\) memory modes. We prove a lower bound \(e_n\) on the probability of a local error (reaching a \(\perp\) state, or seeing a mean payoff \(\leq -1\)) in the current \(n\)-th gadget. This lower bound \(e_n\) holds regardless of events in past gadgets, regardless of the memory mode of \(\sigma\) upon entering the \(n\)-th gadget, and cannot be improved by \(\sigma\) randomizing its memory updates.

The main idea is that, once \(k(n) > k + 1\) (which holds for \(n \geq N'\) sufficiently large) by the Pigeonhole Principle there will always be a memory mode confusing at least two different branches \(i(n), j(n) \neq k(n)\) of the previous random choice at state \(s_n\). This confusion yields a probability \(\geq e_n\) of reaching a \(\perp\) state or seeing a mean payoff \(\leq -1\), regardless of events in past gadgets and regardless of the memory upon entering the \(n\)-th gadget. We show that \(\sum_{n \geq N'} e_n\) is a divergent series. Thus, \(\prod_{n \geq N'}(1 - e_n) = 0\). Hence, \(\mathcal{P}_{M,\sigma,s_0}(\text{MP}_{\lim \inf \geq 0}) \leq \prod_{n \geq N'}(1 - e_n) = 0\).

Lemma 6 and Lemma 7 yield the following theorem.

\textbf{Theorem 8.} There exists a countable, finitely branching and acyclic MDP \(M\) whose step counter is implicit in the state for which \(\text{val}_M,\text{MP}_{\lim \inf \geq 0}(s_0) = 1\) and any FR strategy \(\sigma\) is such that \(\mathcal{P}_{M,s_0,\sigma}(\text{MP}_{\lim \inf \geq 0}) = 0\). In particular, there are no \(\varepsilon\)-optimal \(k\)-bit Markov strategies for any \(k \in \mathbb{N}\) and any \(\varepsilon < 1\) for \(\text{MP}_{\lim \inf \geq 0}\) in countable MDPs.

All of the above results/proofs also hold for \(\text{TP}_{\lim \inf \geq 0}\), giving us the following theorem.

\textbf{Theorem 9.} There exists a countable, finitely branching and acyclic MDP \(M\) whose step counter is implicit in the state for which \(\text{val}_M,\text{TP}_{\lim \inf \geq 0}(s_0) = 1\) and any FR strategy \(\sigma\) is such that \(\mathcal{P}_{M,s_0,\sigma}(\text{TP}_{\lim \inf \geq 0}) = 0\). In particular, there are no \(\varepsilon\)-optimal \(k\)-bit Markov strategies for any \(k \in \mathbb{N}\) and any \(\varepsilon < 1\) for \(\text{TP}_{\lim \inf \geq 0}\) in countable MDPs.

### 3.2 Optimal strategies

Even for acyclic MDPs with the step counter implicit in the state, optimal (and even almost sure winning) strategies for \(\text{MP}_{\lim \inf \geq 0}\) require infinite memory. To prove this, we consider a variant of the MDP from the previous section which has been augmented to include restarts from the \(\perp\) states. For the rest of the section, \(M\) is the MDP constructed in Figure 3.

\textbf{Remark 10.} \(M\) is acyclic, finitely branching and the step counter is implicit in the state. We now refer to the rows of Figure 3 as gadgets, i.e., a gadget is a single instance of Figure 2 where the \(\perp\) states lead to the next row.

\textbf{Lemma 11.} There exists a strategy \(\sigma\) such that \(\mathcal{P}_{M,s_0,\sigma}(\text{MP}_{\lim \inf \geq 0}) = 1\).

\textbf{Proof sketch.} Recall the strategy \(\sigma_{1/2}\) defined in Lemma 6 which achieves at least \(1/2\) in each gadget that it is played in. We then construct the almost surely winning strategy \(\sigma\) by concatenating \(\sigma_{1/2}\) strategies in the sense that \(\sigma\) plays just like \(\sigma_{1/2}\) in each gadget from each gadget’s start state.

Since \(\sigma\) achieves at least \(1/2\) in every gadget that it sees, with probability 1, runs generated by \(\sigma\) restart only finitely many times. The intuition is then that a run restarting finitely many times must spend an infinite tail in some final gadget. Since \(\sigma\) mimics in every controlled state, not restarting anymore directly implies that the total payoff is eventually always \(\geq 0\). Hence all runs generated by \(\sigma\) and restarting only finitely many times satisfy \(\text{MP}_{\lim \inf \geq 0}\). Therefore all but a nullset of runs generated by \(\sigma\) are winning, i.e. \(\mathcal{P}_{M,s_0,\sigma}(\text{MP}_{\lim \inf \geq 0}) = 1\). \qed
Each row represents a copy of the MDP depicted in Figure 2. Each white circle labeled with a number $i$ represents the correspondingly numbered gadget (like in Figure 1) from that MDP. Now, instead of the bottom states in each gadget leading to an infinite losing chain, they lead to a restart state $r_{i,j}$ which leads to a fresh copy of the MDP (in the next row). Each restart incurs a penalty guaranteeing that the mean payoff dips below $-1$ before refunding it and continuing on in the next copy of the MDP. The states $r_{i,j}$ are labeled such that the $j$ indicates that if a run sees this state, then it is the $j$th restart. The $i$ indicates that the run entered the restart state from the $i$th gadget of the current copy of the MDP. The black states are dummy states inserted in order to preserve path length throughout.

Lemma 12. For any FR strategy $\sigma$, $P_{M,\sigma,s_0}(\text{MP lim inf} \geq 0) = 0$.

Proof sketch. Let $\sigma$ be any FR strategy. We partition the runs generated by $\sigma$ into runs restarting infinitely often, and those restarting only finitely many times. Any runs restarting infinitely often are losing by construction. Those runs restarting only finitely many times, once in the gadget they spend an infinite tail in, let the mean payoff dip below $-1$ infinitely many times with probability 1 by Lemma 7. Hence we have that $P_{M,\sigma,s_0}(\text{MP lim inf} \geq 0) = 0$.

From Lemma 11 and Lemma 12 we obtain the following theorem.

Theorem 13. There exists a countable, finitely branching and acyclic MDP $M$ whose step counter is implicit in the state for which $s_0$ is almost surely winning $\text{MP lim inf} \geq 0$, i.e., $\exists \hat{\sigma} P_{M,s_0,\hat{\sigma}}(\text{MP lim inf} \geq 0) = 1$, but every FR strategy $\sigma$ is such that $P_{M,s_0,\sigma}(\text{MP lim inf} \geq 0) = 0$. In particular, almost sure winning strategies, when they exist, cannot be chosen $k$-bit Markov for any $k \in \mathbb{N}$ for countable MDPs.

All of the above results/proofs also hold for $\text{TP lim inf} \geq 0$, giving us the following theorem.

Theorem 14. There exists a countable, finitely branching and acyclic MDP $M$ whose step counter is implicit in the state for which $s_0$ is almost surely winning $\text{TP lim inf} \geq 0$, i.e., $\exists \hat{\sigma} P_{M,s_0,\hat{\sigma}}(\text{TP lim inf} \geq 0) = 1$, but every FR strategy $\sigma$ is such that $P_{M,s_0,\sigma}(\text{TP lim inf} \geq 0) = 0$. In particular, almost sure winning strategies, when they exist, cannot be chosen $k$-bit Markov for any $k \in \mathbb{N}$ for countable MDPs.
When is a reward counter not sufficient?

In this part we show that a reward counter plus arbitrary finite memory does not suffice for \((\varepsilon,\delta,\sigma)\)-optimal strategies for $\text{MP}_{\lim \inf \geq 0}$, even if the MDP is finitely branching.

The same lower bound holds for $\text{TP}_{\lim \inf \geq 0}$/$\text{PP}_{\lim \inf \geq 0}$, but only in infinitely branching MDPs. The finitely branching case is different for $\text{TP}_{\lim \inf \geq 0}$/$\text{PP}_{\lim \inf \geq 0}$; cf. Section 5.

The techniques used to prove these results are similar to those in Section 3 and proofs can be found in [17].

**Theorem 15.** There exists a countable, finitely branching, acyclic MDP $\mathcal{M}_{R1}$ with initial state $(s_0,0)$ with the total reward implicit in the state such that

- $\text{val}_{\mathcal{M}_{R1},\text{MP}_{\lim \inf \geq 0}}((s_0,0)) = 1$,
- for all FR strategies $\sigma$, we have $\mathcal{P}_{\mathcal{M}_{R1},(s_0,0),\sigma}(\text{MP}_{\lim \inf \geq 0}) = 0$.

**Theorem 16.** There exists a countable, finitely branching and acyclic MDP $\mathcal{M}_{Restart}$ whose total reward is implicit in the state where, for the initial state $s_0$,

- there exists an HD strategy $\sigma$ s.t. $\mathcal{P}_{\mathcal{M}_{Restart},s_0,\sigma}(\text{MP}_{\lim \inf \geq 0}) = 1$.
- for every FR strategy $\sigma$, $\mathcal{P}_{\mathcal{M}_{Restart},s_0,\sigma}(\text{MP}_{\lim \inf \geq 0}) = 0$.

**Theorem 17.** There exists an infinitely branching MDP $\mathcal{M}$ with reward implicit in the state and initial state $s$ such that

- every FR strategy $\sigma$ is such that $\mathcal{P}_{\mathcal{M},s,\sigma}(\text{TP}_{\lim \inf \geq 0}) = 0$ and $\mathcal{P}_{\mathcal{M},s,\sigma}(\text{PP}_{\lim \inf \geq 0}) = 0$.
- there exists an HD strategy $\sigma$ s.t. $\mathcal{P}_{\mathcal{M},s,\sigma}(\text{TP}_{\lim \inf \geq 0}) = 1$ and $\mathcal{P}_{\mathcal{M},s,\sigma}(\text{PP}_{\lim \inf \geq 0}) = 1$.

Hence, optimal (and even almost-surely winning) strategies and $\varepsilon$-optimal strategies for $\text{TP}_{\lim \inf \geq 0}$ and $\text{PP}_{\lim \inf \geq 0}$ require infinite memory beyond a reward counter.

**Remark 18.** The MDPs from Section 3 and Section 4 show that good strategies for $\text{MP}_{\lim \inf \geq 0}$ require at least (in the sense of Remark 1) a reward counter and a step counter, respectively. There does, of course, exist a single MDP where good strategies for $\text{MP}_{\lim \inf \geq 0}$ require at least both a step counter and a reward counter. We construct such an MDP by “gluing” the two different MDPs together via an initial random state which points to each with probability $1/2$.

5 Upper bounds

We establish upper bounds on the strategy complexity of $\lim \inf$ threshold objectives for mean payoff, total payoff and point payoff. It is noteworthy that once the reward structure of an MDP has been encoded into the states, then these threshold objectives take on a qualitative flavor not dissimilar to Safety or co-Büchi (cf. [16]). Indeed, if the transition rewards are restricted to integer values, then $\text{TP}_{\lim \inf \geq 0}$ boils down to eventually avoiding all transitions with negative reward (since negative rewards would be $\leq -1$). This is a co-Büchi objective. However, if the rewards are not restricted to integers, then the picture is not so simple.

For finitely branching MDPs, we show that there exist $\varepsilon$-optimal MD strategies for $\text{PP}_{\lim \inf \geq 0}$. In turn, this yields the requisite upper bound for finitely branching $\text{TP}_{\lim \inf \geq 0}$, i.e., using just a reward counter.

For infinitely branching MDPs, a step counter suffices in order to achieve $\text{PP}_{\lim \inf \geq 0}$ $\varepsilon$-optimally. Then, by encoding the total reward into the states, this will also give us SC+RC upper bounds for $\text{MP}_{\lim \inf \geq 0}$ as well as infinitely branching $\text{TP}_{\lim \inf \geq 0}$ (i.e., using both a step counter and a reward counter).

First we show how to encode the total reward level into the state in a given MDP.
Remark 19. Given an MDP $\mathcal{M}$ and initial state $s_0$, we can construct an MDP $R(\mathcal{M})$ with initial state $(s_0, 0)$ and with the reward counter implicit in the state such that strategies in $R(\mathcal{M})$ can be translated back to $\mathcal{M}$ with an extra reward counter.

By labeling transitions in $R(\mathcal{M})$ with the state encoded total reward of the target state, we ensure that the point rewards in $R(\mathcal{M})$ correspond exactly to the total rewards in $\mathcal{M}$.

Lemma 20. Let $\mathcal{M}$ be an MDP with initial state $s_0$. Then given an MD (resp. Markov) strategy $\sigma'$ in $R(\mathcal{M})$ attaining $c \in [0, 1]$ for $PP_{\lim_{\inf \tau \geq 0}}$ from $(s_0, 0)$, there exists a strategy $\sigma$ attaining $c$ for $TP_{\lim_{\inf \tau \geq 0}}$ in $\mathcal{M}$ from $s_0$ which uses the same memory as $\sigma'$ plus a reward counter.

Remark 21. Given an MDP $\mathcal{M}$ and initial state $s_0$, we can construct an acyclic MDP $A(\mathcal{M})$ with initial state $(s_0, 0)$ and with the step counter implicit in the state such that MD strategies in $A(\mathcal{M})$ can be translated back to $\mathcal{M}$ with the use of a step counter to yield deterministic Markov strategies in $\mathcal{M}$; cf. [15, Lemma 4].

Remark 22. In order to tackle the mean payoff objective $MP_{\lim_{\inf \tau \geq 0}}$ on $\mathcal{M}$, we define a new acyclic MDP $A(\mathcal{M})$ which encodes both the step counter and the average reward into the state. However, since we want the point rewards in $A(\mathcal{M})$ to coincide with the mean payoff in the original MDP $\mathcal{M}$, the transition rewards in $A(\mathcal{M})$ are given as the encoded rewards divided by the step counter (unlike in $R(\mathcal{M})$).

Lemma 23. Let $\mathcal{M}$ be an MDP with initial state $s_0$. Then given an MD strategy $\sigma'$ in $A(\mathcal{M})$ attaining $c \in [0, 1]$ for $PP_{\lim_{\inf \tau \geq 0}}$ from $(s_0, 0, 0)$, there exists a strategy $\sigma$ attaining $c$ for $MP_{\lim_{\inf \tau \geq 0}}$ in $\mathcal{M}$ from $s_0$ which uses just a reward counter and a step counter.

Proof. The proof is very similar to that of Lemma 20.

Lemma 24 ([15, Lemma 23]). For every acyclic MDP with a safety objective and every $\varepsilon > 0$, there exists an MD strategy that is uniformly $\varepsilon$-optimal.

Theorem 25 ([13, Theorem 7]). Let $\mathcal{M} = (S, S_0, S_\text{safe}, \to, P, r)$ be a countable MDP, and let $\varphi$ be an event that is tail in $\mathcal{M}$. Suppose for every $s \in S$ there exist $\varepsilon$-optimal MD strategies for $\varphi$. Then:
1. There exist uniform $\varepsilon$-optimal MD strategies for $\varphi$.
2. There exists a single MD strategy that is optimal from every state that has an optimal strategy.

5.1 Finitely Branching Case

In order to prove the main result of this section, we use the following result on the Transience objective, which is the set of runs that do not visit any state infinitely often. Given an MDP $\mathcal{M} = (S, S_0, S_\text{safe}, \to, P, r)$, $\text{Transience} = \bigwedge_{s \in S} \text{FG}\neg s$.

Theorem 26 ([13, Theorem 8]). In every countable MDP there exist uniform $\varepsilon$-optimal MD strategies for Transience.

Theorem 27. Consider a finitely branching MDP $\mathcal{M} = (S, S_0, S_\text{safe}, \to, P, r)$ with initial state $s_0$ and a $PP_{\lim_{\inf \tau \geq 0}}$ objective. Then there exist $\varepsilon$-optimal MD strategies.

Proof. Let $\varepsilon > 0$. We begin by partitioning the state space into two sets, $S_\text{safe}$ and $S \setminus S_\text{safe}$. The set $S_\text{safe}$ is the subset of states which is surely winning for the safety objective of only using transitions with non-negative rewards (i.e., never using transitions with negative rewards at all). Since $\mathcal{M}$ is finitely branching, there exists a uniformly optimal MD strategy $\sigma_\text{safe}$ for this safety objective [18, 16].
12:12 Strategy Complexity of Mean/Total/Point Payoff Objectives in Countable MDPs

We construct a new MDP $\mathcal{M}'$ by modifying $\mathcal{M}$. We create a gadget $G_{\text{safe}}$ composed of a sequence of new controlled states $x_0, x_1, x_2, \ldots$ where all transitions $x_i \rightarrow x_{i+1}$ have reward 0. Hence any run entering $G_{\text{safe}}$ is winning for $PP_{\lim \inf \geq 0}$. We insert $G_{\text{safe}}$ into $\mathcal{M}$ by replacing all incoming transitions to $S_{\text{safe}}$ with transitions that lead to $x_0$. The idea behind this construction is that when playing in $\mathcal{M}$, once you hit a state in $S_{\text{safe}}$, you can win surely by playing an optimal MD strategy for safety. So we replace $S_{\text{safe}}$ with the surely winning gadget $G_{\text{safe}}$. Thus

$$\text{val}_{\mathcal{M},PP_{\lim \inf \geq 0}}(s_0) = \text{val}_{\mathcal{M}',PP_{\lim \inf \geq 0}}(s_0)$$

(1)

and if an $\varepsilon$-optimal MD strategy exists in $\mathcal{M}$, then there exists a corresponding one in $\mathcal{M}'$, and vice-versa.

We now consider a general (not necessarily MD) $\varepsilon$-optimal strategy $\sigma$ for $PP_{\lim \inf \geq 0}$ from $s_0$ on $\mathcal{M}'$, i.e.,

$$\mathcal{P}_{\mathcal{M}',s_0,\sigma}(PP_{\lim \inf \geq 0}) \geq \text{val}_{\mathcal{M}',PP_{\lim \inf \geq 0}}(s_0) - \varepsilon.$$ 

(2)

Define the safety objective $\text{Safety}_i$ which is the objective of never seeing any point rewards $<-2^{-i}$. This then allows us to characterize $PP_{\lim \inf \geq 0}$ in terms of safety objectives.

$$PP_{\lim \inf \geq 0} = \bigcap_{i \in \mathbb{N}} F(\text{Safety}_i).$$

(3)

Now we define the safety objective $\text{Safety}_i^k \overset{\text{def}}{=} F^{\leq k}(\text{Safety}_i)$ to attain $\text{Safety}_i$ within at most $k$ steps. This allows us to write

$$F(\text{Safety}_i) = \bigcup_{k \in \mathbb{N}} \text{Safety}_i^k.$$ 

(4)

By continuity of measures from above we get

$$0 = \mathcal{P}_{\mathcal{M}',s_0,\sigma} \left( F(\text{Safety}_i) \cap \bigcap_{k \in \mathbb{N}} \text{Safety}_i^k \right) = \lim_{k \rightarrow \infty} \mathcal{P}_{\mathcal{M}',s_0,\sigma} \left( F(\text{Safety}_i) \cap \overline{\text{Safety}_i^k} \right).$$

Hence for every $i \in \mathbb{N}$ and $\varepsilon < 2^{-i}$ there exists $n_i$ such that

$$\mathcal{P}_{\mathcal{M}',s_0,\sigma} \left( F(\text{Safety}_i) \cap \overline{\text{Safety}_i^{n_i}} \right) \leq \varepsilon.$$ 

(5)

Now we can show the following claim (proof in [17]).

Claim 28.

$$\mathcal{P}_{\mathcal{M}',s_0,\sigma} \left( \bigcap_{i \in \mathbb{N}} \text{Safety}_i^{n_i} \right) \geq \text{val}_{\mathcal{M}',PP_{\lim \inf \geq 0}}(s_0) - 2\varepsilon.$$ 

Since $\mathcal{M}'$ does not have an implicit step counter, we use the following construction to approximate one. We define the distance $d(s)$ from $s_0$ to a state $s$ as the length of the shortest path from $s_0$ to $s$. Let $\text{Bubble}_{n}(s_0) \overset{\text{def}}{=} \{ s \in S | d(s) \leq n \}$ be those states that can be reached within $n$ steps from $s_0$. Since $\mathcal{M}'$ is finitely branching, $\text{Bubble}_{n}(s_0)$ is finite for every fixed $n$. Let

$$\text{Bad}_i \overset{\text{def}}{=} \{ t \in \rightarrow^* \mathcal{M}' | t = s \rightarrow^* \mathcal{M}' s', s \notin \text{Bubble}_{n_i}(s_0) \text{ and } r(t) < -2^{-i} \}.$$
be the set of transitions originating outside Bubble\(i\)\((s_0)\) whose reward is too negative. Thus a run from \(s_0\) that satisfies Safety\(i\) cannot use any transition in Bad\(i\), since (by definition of Bubble\(i\)\((s_0)\)) they would come after the \(n_i\)-th step.

Now we create a new state \(\perp\) whose only outgoing transition is a self loop with reward \(-1\). We transform \(M'\) into \(M''\) by re-directing all transitions in Bad\(i\), to the new target state \(\perp\) for every \(i\). Notice that any run visiting \(\perp\) must be losing for \(PP_{\lim \inf \geq 0}\) due to the negative reward on the self loop, but it must also be losing for Transience because of the self loop.

We now show that the change from \(M'\) to \(M''\) has decreased the value of \(s_0\) for \(PP_{\lim \inf \geq 0}\) by at most \(2\varepsilon\), i.e.,

\[
\val_{M'', PP_{\lim \inf \geq 0}}(s_0) \geq \val_{M', PP_{\lim \inf \geq 0}}(s_0) - 2\varepsilon.
\]  

(6)

Equation (6) follows from the following steps.

\[
\val_{M'', PP_{\lim \inf \geq 0}}(s_0) \geq \mathcal{P}_{M'', s_0, \sigma} \left( \bigcap_{i \in \mathbb{N}} \text{Safety}_{i}^{n_i} \right) \\
= \mathcal{P}_{M', s_0, \sigma} \left( \bigcap_{i \in \mathbb{N}} \text{Safety}_{i}^{n_i} \right) \quad \text{by def. of } M'' \\
\geq \val_{M', PP_{\lim \inf \geq 0}}(s_0) - 2\varepsilon \quad \text{by Claim 28}
\]

In the next step (proof in [17]) we argue that under every strategy \(\sigma''\) from \(s_0\) in \(M''\) the attainment for \(PP_{\lim \inf \geq 0}\) and Transience coincide, i.e.,

\[\forall \sigma''. \mathcal{P}_{M'', s_0, \sigma''}(PP_{\lim \inf \geq 0}) = \mathcal{P}_{M'', s_0, \sigma''}(\text{Transience}).\]

By Theorem 26, there exists a uniformly \(\varepsilon\)-optimal MD strategy \(\hat{\sigma}\) from \(s_0\) for Transience in \(M''\), i.e.,

\[\mathcal{P}_{M'', s_0, \hat{\sigma}}(\text{Transience}) \geq \val_{M'', \text{Transience}}(s_0) - \varepsilon.\]  

(7)

We construct an MD strategy \(\sigma^*\) in \(M\) which plays like \(\sigma_{\text{safe}}\) in \(S_{\text{safe}}\) and plays like \(\hat{\sigma}\) everywhere else.

\[
\mathcal{P}_{M, s_0, \sigma^*}(PP_{\lim \inf \geq 0}) = \mathcal{P}_{M', s_0, \hat{\sigma}}(PP_{\lim \inf \geq 0}) \quad \text{def. of } \sigma^* \text{ and } \sigma_{\text{safe}} \\
\geq \mathcal{P}_{M', s_0, \hat{\sigma}}(PP_{\lim \inf \geq 0}) \quad \text{new losing sink in } M'' \\
= \mathcal{P}_{M', s_0, \hat{\sigma}}(\text{Transience}) \quad \text{by Claim 29} \\
\geq \val_{M', \text{Transience}}(s_0) - \varepsilon \quad \text{by (7)} \\
= \val_{M'', PP_{\lim \inf \geq 0}}(s_0) - \varepsilon \quad \text{by Claim 29} \\
\geq \val_{M', PP_{\lim \inf \geq 0}}(s_0) - 2\varepsilon - \varepsilon \quad \text{by (6)} \\
= \val_{M, PP_{\lim \inf \geq 0}}(s_0) - 3\varepsilon \quad \text{by (1)}
\]

Hence \(\sigma^*\) is a \(3\varepsilon\)-optimal MD strategy for \(PP_{\lim \inf \geq 0}\) from \(s_0\) in \(M\) as required.

\[\blacktriangleright\text{Corollary 30.} \text{ Given a finitely branching MDP } M, \text{ there exist } \varepsilon\text{-optimal strategies for } TP_{\lim \inf \geq 0} \text{ which use just a reward counter.}\]

\[\blacktriangleright\text{Proof.} \text{ By Theorem 27 and Lemma 20.}\]

\[\blacktriangleright\]
Corollary 31. Given a finitely branching MDP $M$ and initial state $s_0$, optimal strategies, where they exist,

- for $PP_{\lim\inf \geq 0}$ can be chosen MD.
- for $TP_{\lim\inf \geq 0}$ can be chosen with just a reward counter.

Proof. Since $PP_{\lim\inf \geq 0}$ is tail, the first claim follows from Theorem 27 and Theorem 25. Towards the second claim, we place ourselves in $R(M)$ where $TP_{\lim\inf \geq 0}$ is tail. Moreover, in $R(M)$ the objectives $TP_{\lim\inf \geq 0}$ and $PP_{\lim\inf \geq 0}$ coincide. Thus we can apply Theorem 27 to obtain $\varepsilon$-optimal MD strategies for $TP_{\lim\inf \geq 0}$ from every state of $R(M)$. From Theorem 25 we obtain a single MD strategy that is optimal from every state of $R(M)$ that has an optimal strategy. By Lemma 20 we can translate this MD strategy on $R(M)$ back to a strategy on $M$ with just a reward counter. ▶

5.2 Infinitely Branching Case

For infinitely branching MDPs, $\varepsilon$-optimal strategies for $PP_{\lim\inf \geq 0}$ require more memory than in the finitely branching case. However, the proofs are similar to those in Section 5.1 and can be found in [17].

Theorem 32. Consider an MDP $M$ with initial state $s_0$ and a $PP_{\lim\inf \geq 0}$ objective. For every $\varepsilon > 0$ there exist

- $\varepsilon$-optimal MD strategies in $S(M)$.
- $\varepsilon$-optimal deterministic Markov strategies in $M$.

Corollary 33. Given an MDP $M$ and initial state $s_0$, there exist $\varepsilon$-optimal strategies $\sigma$ for $MP_{\lim\inf \geq 0}$ which use just a step counter and a reward counter.

Corollary 34. Given an MDP $M$ with initial state $s_0$,

- there exist $\varepsilon$-optimal MD strategies for $TP_{\lim\inf \geq 0}$ in $S(R(M))$,
- there exist $\varepsilon$-optimal strategies for $TP_{\lim\inf \geq 0}$ which use a step counter and a reward counter.

Corollary 35. Given an MDP $M$ and initial state $s_0$, optimal strategies, where they exist,

- for $PP_{\lim\inf \geq 0}$ can be chosen with just a step counter.
- for $MP_{\lim\inf \geq 0}$ and $TP_{\lim\inf \geq 0}$ can be chosen with just a reward counter and a step counter.

6 Conclusion and Outlook

We have established matching lower and upper bounds on the strategy complexity of $\lim\inf$ threshold objectives for point, total and mean payoff on countably infinite MDPs; cf. Table 1.

The upper bounds hold not only for integer transition rewards, but also for rationals or reals, provided that the reward counter (in those cases where one is required) is of the same type. The lower bounds hold even for integer transition rewards, since all our counterexamples are of this form.

Directions for future work include the corresponding questions for $\lim\sup$ threshold objectives. While the $\lim\inf$ point payoff objective generalizes co-Büchi (see Section 2), the $\lim\sup$ point payoff objective generalizes Büchi. Thus the lower bounds for $\lim\sup$ point payoff are at least as high as the lower bounds for Büchi objectives [14, 15].
References


