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

Citation for published version:
Mayr, R & Munday, E 2021, Strategy Complexity of Mean Payoff, Total Payoff and Point Payoff Objectives in Countable MDPs. in S Haddad & D Varacca (eds), *32nd International Conference on Concurrency Theory (CONCUR 2021).* LIPIcs - Leibniz International Proceedings in Informatics, vol. 203, Schloss Dagstuhl - Leibniz-Zentrum für Informatik, pp. 12:1-12:15, 32nd International Conference on Concurrency Theory, Paris, France, 23/08/21. https://doi.org/10.4230/LIPIcs.CONCUR.2021.12

Digital Object Identifier (DOI):
10.4230/LIPIcs.CONCUR.2021.12

Published In:
32nd International Conference on Concurrency Theory (CONCUR 2021)

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 $\liminf$ 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 ε-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

Digital Object Identifier 10.4230/LIPIcs.CONCUR.2021.12

Related Version *Full Version*: `https://arxiv.org/abs/2107.03287`

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 ε-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 $r_0 r_1 r_2 \dots$. We consider the $\liminf$ of this sequence, as well as two other important derived sequences.

1. The point payoff considers the $\liminf$ of the sequence $r_0 r_1 r_2 \dots$ directly.
2. The total payoff considers the $\liminf$ 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 $\liminf$ 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 $\liminf$ threshold objective is to maximize the probability that the $\liminf$ of the respective type of sequence is ≥ 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
LIPICS 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 1 Strategy complexity of ε-optimal/optimal strategies for point, total and mean payoff objectives in infinitely/finitely branching MDPs. MD stands for memoryless deterministic, SC for step counter, RC for reward counter and SC+RC for both. All strategies are deterministic and randomization does not help. For each result, we list the numbers of the theorems that show the upper and lower bounds on the strategy complexity. The lower bounds hold in the sense of Remark 1, but work for integer rewards. The upper bounds hold even for real-valued rewards.

	Point payoff	Total payoff	Mean payoff
ε-optimal, infinitely branching	SC 17, 32	SC+RC 17, 9, 34	SC+RC 15, 8, 33
optimal, infinitely branching	SC 17, 35	SC+RC 14, 17, 35	SC+RC 13, 16, 35
ε-optimal, finitely branching	MD 27	RC 9, 30	SC+RC 15, 8, 33
optimal, finitely branching	MD 31	RC 14, 31	SC+RC 13, 16, 35

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 $\liminf \geq 0$, 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), (ε-)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 $\liminf$ mean payoff need not coincide with attainment for $\limsup$ mean payoff, even for very simple examples. E.g., consider the acyclic infinite graph with transitions $s_n \to s_{n+1}$ for all $n \in \mathbb{N}$ with reward $(-1)^n 2^n$ in the n-th step, which yields a $\liminf$ mean payoff of $-\infty$ and a $\limsup$ 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 MD (memoryless deterministic) strategies for $\liminf/\limsup$ mean payoff. [19, Counterexample 1.3] shows that there are not even ε-optimal memoryless randomized strategies for $\liminf/\limsup$ 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 $\liminf/\limsup$ is non-negative, it asks to maximize the *expectation* of the $\liminf/\limsup$ point payoffs, which is a different problem (e.g., it can tolerate a high probability of a negative $\liminf/\limsup$ if the remaining cases have a huge positive $\liminf/\limsup$). Hill & Pestien [12] showed the existence of good randomized Markov strategies for the $\limsup$ of the *expected* average reward up-to step n for growing n, and for the *expected* $\liminf$ of the point payoffs.

2 Preliminaries

Markov decision processes. A *probability distribution* over a countable set S is a function $f : S \to [0, 1]$ 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) $\mathcal{M} = (S, S_\Box, S_\bigcirc, {\longrightarrow}, P, r)$ consists of a countable set S of *states*, which is partitioned into a set $S_\Box$ of *controlled states* and a set $S_\bigcirc$ of *random states*, a *transition relation* ${\longrightarrow} \subseteq S \times S$, and a *probability function* $P : S_\bigcirc \to \mathcal{D}(S)$. We write $s{\longrightarrow}s'$ if $(s, s') \in {\longrightarrow}$, 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 $s \in S_\bigcirc$ 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 $\longrightarrow$ relation, that is, $s \in T$ and $s{\longrightarrow}s'$ implies that $s' \in T$.

An MDP is *acyclic* if the underlying directed graph $(S, {\longrightarrow})$ 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 ($S_\Box = \emptyset$) is called a *Markov chain*.

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

Strategies and Probability Measures. A *run* ρ is an infinite sequence of states and transitions $s_0 e_0 s_1 e_1 \cdots$ such that $e_i = (s_i, s_{i+1}) \in {\longrightarrow}$ for all $i \in \mathbb{N}$. Let $\mathit{Runs}_{\mathcal{M}}^{s_0}$ be the set of all runs from s_0 in the MDP $\mathcal{M}$. A *partial run* is a finite prefix of a run, $\mathit{pRuns}_{\mathcal{M}}^{s_0}$ is the set of all partial runs from s_0 and $\mathit{pRuns}_{\mathcal{M}}$ the set of partial runs from any state.

We write $\rho_s(i) \overset{\text{def}}{=} s_i$ for the i-th state along ρ and $\rho_e(i) \overset{\text{def}}{=} e_i$ for the i-th transition along ρ. We sometimes write runs as $s_0 s_1 \cdots$, leaving the transitions implicit. We say that a (partial) run ρ *visits* s if $s = \rho_s(i)$ for some i, and that ρ starts in s if $s = \rho_s(0)$.

A *strategy* is a function $\sigma : \mathit{pRuns}_{\mathcal{M}} \cdot S_\Box \to \mathcal{D}(S)$ that assigns to partial runs ρs, where $s \in S_\Box$, a distribution over the successors $\{s' \in S \mid s{\longrightarrow}s'\}$. The set of all strategies in $\mathcal{M}$ is denoted by $\Sigma_{\mathcal{M}}$ (we omit the subscript and write Σ if $\mathcal{M}$ is clear from the context). A (partial) run $s_0 e_0 s_1 e_1 \cdots$ is consistent with a strategy σ if for all i either $s_i \in S_\Box$ and $\sigma(s_0 e_0 s_1 e_1 \cdots s_i)(s_{i+1}) > 0$, or $s_i \in S_\bigcirc$ and $P(s_i)(s_{i+1}) > 0$.

An MDP $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P, r)$, an initial state $s_0 \in S$, and a strategy σ induce a probability space in which the outcomes are runs starting in s_0 and with measure $\mathcal{P}_{\mathcal{M},s_0,\sigma}$ defined as follows. It is first defined on *cylinders* $s_0 e_0 s_1 e_1 \ldots s_n Runs^{s_n}_{\mathcal{M}}$: if $s_0 e_0 s_1 e_1 \ldots s_n$ is not a partial run consistent with σ then $\mathcal{P}_{\mathcal{M},s_0,\sigma}(s_0 e_0 s_1 e_1 \ldots s_n Runs^{s_n}_{\mathcal{M}}) \stackrel{\text{def}}{=} 0$. Otherwise, $\mathcal{P}_{\mathcal{M},s_0,\sigma}(s_0 e_0 s_1 e_1 \ldots s_n Runs^{s_n}_{\mathcal{M}}) \stackrel{\text{def}}{=} \prod_{i=0}^{n-1} \bar{\sigma}(s_0 e_0 s_1 \ldots s_i)(s_{i+1})$, where $\bar{\sigma}$ is the map that extends σ by $\bar{\sigma}(ws) = P(s)$ for all partial runs $ws \in pRuns_{\mathcal{M}} \cdot S_\bigcirc$. By Carathéodory's theorem [4], this extends uniquely to a probability measure $\mathcal{P}_{\mathcal{M},s_0,\sigma}$ on the Borel σ-algebra $\mathcal{F}$ of subsets of $Runs^{s_0}_{\mathcal{M}}$. Elements of $\mathcal{F}$, i.e., measurable sets of runs, are called *events* or *objectives* here. For $X \in \mathcal{F}$ we will write $\overline{X} \stackrel{\text{def}}{=} Runs^{s_0}_{\mathcal{M}} \setminus X \in \mathcal{F}$ for its complement and $\mathcal{E}_{\mathcal{M},s_0,\sigma}$ for the expectation wrt. $\mathcal{P}_{\mathcal{M},s_0,\sigma}$. 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 $(S, \longrightarrow)$. We use $[\![\varphi]\!]^s$ to denote the set of runs starting from s that satisfy the LTL formula φ, 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 φ and $[\![\varphi]\!]$ and just write $\mathcal{P}_{\mathcal{M},s,\sigma}(\varphi)$ instead of $\mathcal{P}_{\mathcal{M},s,\sigma}([\![\varphi]\!]^s)$. The reachability objective of eventually visiting a set of states X can be expressed by $[\![\mathsf{F} X]\!] \stackrel{\text{def}}{=} \{\rho \mid \exists i.\, \rho_s(i) \in X\}$. Reaching X within at most k steps is expressed by $[\![\mathsf{F}^{\leq k} X]\!] \stackrel{\text{def}}{=} \{\rho \mid \exists i \leq k.\, \rho_s(i) \in X\}$. The definitions for eventually visiting certain transitions are analogous. The operator G (always) is defined as $\neg\mathsf{F}\neg$. So the safety objective of avoiding X is expressed by $\mathsf{G}\neg X$.

- The $PP_{\liminf \geq 0}$ objective is to maximize the probability that the lim inf of the *point* payoffs (the immediate transition rewards) is ≥ 0, i.e., $PP_{\liminf \geq 0} \stackrel{\text{def}}{=} \{\rho \mid \liminf_{n \in \mathbb{N}} r(\rho_e(n)) \geq 0\}$.
- 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 ≥ 0, i.e., $TP_{\liminf \geq 0} \stackrel{\text{def}}{=} \{\rho \mid \liminf_{n \in \mathbb{N}} \sum_{j=0}^{n-1} r(\rho_e(j)) \geq 0\}$.
- The $MP_{\liminf \geq 0}$ objective is to maximize the probability that the lim inf of the *mean* payoff is ≥ 0, i.e., $MP_{\liminf \geq 0} \stackrel{\text{def}}{=} \{\rho \mid \liminf_{n \in \mathbb{N}} \frac{1}{n} \sum_{j=0}^{n-1} r(\rho_e(j)) \geq 0\}$.

An objective φ is called *tail* in $\mathcal{M}$ if for every run $\rho'\rho$ in $\mathcal{M}$ with some finite prefix ρ' 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 ≤ -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 $\mathcal{D}(S)$. A strategy σ is *deterministic* (D) if $\sigma(\rho)$ is a Dirac distribution for all ρ. 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 σ is *memoryless* (M) (also called *positional*) if it can be implemented with a memory of size 1. We may view M-strategies as functions $\sigma : S_\Box \to \mathcal{D}(S)$.
- A strategy σ is *finite memory* (F) if there exists a finite memory M implementing σ. 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 to 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 ε-optimal Strategies. Given an objective φ, the value of state s in an MDP $\mathcal{M}$, denoted by $\mathtt{val}_{\mathcal{M},\varphi}(s)$, is the supremum probability of achieving φ. Formally, $\mathtt{val}_{\mathcal{M},\varphi}(s) \stackrel{\text{def}}{=} \sup_{\sigma\in\Sigma} \mathcal{P}_{\mathcal{M},s,\sigma}(\varphi)$ where Σ is the set of all strategies. For $\varepsilon \geq 0$ and state $s \in S$, we say that a strategy is *ε-optimal* from s if $\mathcal{P}_{\mathcal{M},s,\sigma}(\varphi) \geq \mathtt{val}_{\mathcal{M},\varphi}(s) - \varepsilon$. A 0-optimal strategy is called *optimal*. An optimal strategy is *almost-surely winning* if $\mathtt{val}_{\mathcal{M},\varphi}(s) = 1$. Considering an MD strategy as a function $\sigma : S_\Box \to S$ and $\varepsilon \geq 0$, σ is *uniformly* ε-optimal (resp. uniformly optimal) if it is ε-optimal (resp. optimal) from *every* $s \in S$.

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

Lower bounds on the strategy complexity of an objective φ can only be established in the sense of proving that good strategies for φ 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 φ 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 ε-optimal strategies for $MP_{\liminf\geq 0}$ or $TP_{\liminf\geq 0}$. We will construct an acyclic MDP where the step counter is implicit in the state such that ε-optimal strategies for $MP_{\liminf\geq 0}$ and $TP_{\liminf\geq 0}$ still require infinite memory.

3.1 Epsilon-optimal strategies

We construct an acyclic MDP $\mathcal{M}$ 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 $n = N^*$. In the controlled choice, there is a small chance in all but the top choice of falling into a $\bot$ state. These $\bot$ 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.

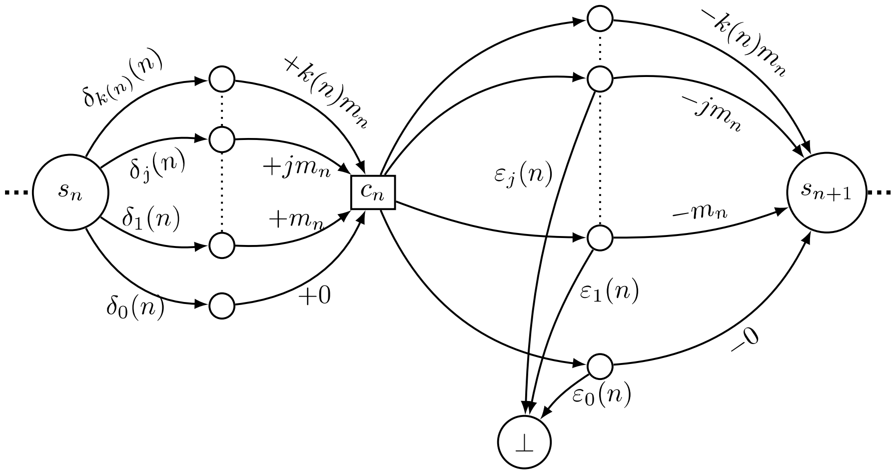

Figure 1 A typical building block with $k(n)+1$ choices, first random then controlled. The number of choices $k(n)+1$ grows unboundedly with n. This is the n-th building block of the MDP in Figure 2. The $\delta_i(n)$ and $\varepsilon_i(n)$ are probabilities depending on n and the $\pm im_n$ are transition rewards. We index the successor states of s_n and c_n from 0 to $k(n)$ to match the indexing of the δ's and ε's such that the bottom state is indexed with 0 and the top state with $k(n)$.

▶ Remark 2. $\mathcal{M}$ 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 ε-optimal strategies.

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

$$\log_1 n \overset{\text{def}}{=} \log n, \quad \log_{i+1} n \overset{\text{def}}{=} \log(\log_i n)$$

$$\delta_0(n) \overset{\text{def}}{=} \frac{1}{\log n}, \quad \delta_i(n) \overset{\text{def}}{=} \frac{1}{\log_{i+1} n}, \quad \delta_{k(n)}(n) \overset{\text{def}}{=} 1 - \sum_{j=0}^{k(n)-1} \delta_j(n)$$

$$\varepsilon_0(n) \overset{\text{def}}{=} \frac{1}{n\log n}, \ \varepsilon_{i+1}(n) \overset{\text{def}}{=} \frac{\varepsilon_i(n)}{\log_{i+2} n}, \text{ i.e. } \varepsilon_i(n) = \frac{1}{n \cdot \log n \cdot \log_2 n \cdots \log_{i+1} n}, \varepsilon_{k(n)}(n) \overset{\text{def}}{=} 0$$

$$\text{Tower}(0) \overset{\text{def}}{=} e^0 = 1, \quad \text{Tower}(i+1) \overset{\text{def}}{=} e^{\text{Tower}(i)}, \quad N_i \overset{\text{def}}{=} \text{Tower}(i)$$

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

$$g(i) \overset{def}{=} min\left\{N : \left(\sum_{n>N} \delta_{i-1}(n)\varepsilon_{i-1}(n)\right) \leq 2^{-i}\right\}, \quad h(1) \overset{def}{=} 2$$

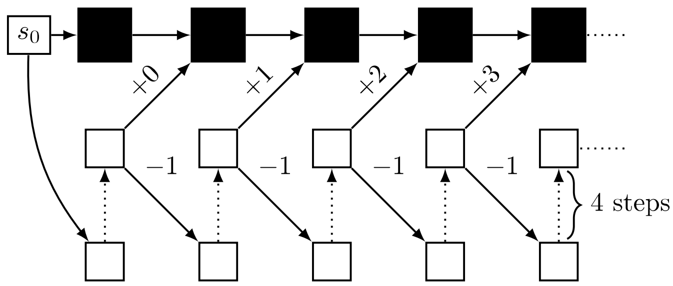

Figure 2 The buildings blocks from Figure 1 represented by black boxes are chained together (n increases as you go to the right). The chain of white boxes allows to skip arbitrarily long prefixes while preserving path length. The positive rewards from the white states to the black boxes reimburse the lost reward accumulated until then. The -1 rewards between white states ensure that skipping gadgets forever is losing.

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

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 $k(n) \stackrel{\text{def}}{=} h^{-1}(n)$. 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: $N^* \stackrel{\text{def}}{=} \mathit{min}\{n \in \mathbb{N} : k(n) = 1\}$. *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 $m_n \stackrel{\text{def}}{=} 2k(n)\sum_{i=N^*}^{n-1} m_i$, with $m_{N^*} \stackrel{\text{def}}{=} 1$ and where $k(n)$ is the branching degree.

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

▶ **Lemma 5.** *For $k(n) \geq 1$, the transition probabilities in the gadgets are well defined.*

▶ **Lemma 6.** *For every $\varepsilon > 0$, there exists a strategy σ_ε with $\mathcal{P}_{\mathcal{M},s_0,\sigma_\varepsilon}(MP_{\liminf \geq 0}) \geq 1-\varepsilon$ that cannot fail unless it hits a $\perp$ state. Formally, $\mathcal{P}_{\mathcal{M},s_0,\sigma_\varepsilon}(MP_{\liminf\geq 0} \wedge \mathsf{G}(\neg \perp)) = \mathcal{P}_{\mathcal{M},s_0,\sigma_\varepsilon}(\mathsf{G}(\neg \perp)) \geq 1-\varepsilon$. So in particular, $\mathtt{val}_{\mathcal{M},MP_{\liminf\geq 0}}(s_0) = 1$.*

Proof sketch. We define a strategy σ which in c_n always mimics the choice in s_n. Playing according to σ, the only way to lose is by dropping into the $\perp$ 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 σ.

When playing with the ability to skip gadgets, as illustrated in Figure 2, all runs not visiting a $\perp$ state are winning since the total reward never dips below 0. We then consider the strategy σ_ε which plays like σ after skipping forwards by sufficiently many gadgets (starting at $n \gg N^*$). Its probability of satisfying $MP_{\liminf\geq 0}$ corresponds to a tail of the above product, which can be made arbitrarily close to 1 (and thus $\geq 1-\varepsilon$). Thus the strategies σ_ε for arbitrarily small $\varepsilon > 0$ witness that $\mathtt{val}_{\mathcal{M},MP_{\liminf\geq 0}}(s_0) = 1$. ◀

▶ **Lemma 7.** *For any FR strategy σ, almost surely either the mean payoff dips below -1 infinitely often, or the run hits a $\bot$ state, i.e. $\mathcal{P}_{\mathcal{M},\sigma,s_0}(MP_{\liminf\geq 0}) = 0$.*

Proof sketch. Let σ be some FR strategy with k memory modes. We prove a *lower bound* e_n on the probability of a local error (reaching a $\bot$ state, or seeing a mean payoff ≤ -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 σ upon entering the n-th gadget, and cannot be improved by σ 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 $\bot$ state or seeing a mean payoff ≤ -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, $\mathcal{P}_{\mathcal{M},\sigma,s_0}(MP_{\liminf\geq 0}) \leq \prod_{n\geq N'}(1-e_n) = 0$. ◀

Lemma 6 and Lemma 7 yield the following theorem.

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

All of the above results/proofs also hold for $TP_{\liminf\geq 0}$, giving us the following theorem.

▶ **Theorem 9.** *There exists a countable, finitely branching and acyclic MDP $\mathcal{M}$ whose step counter is implicit in the state for which $\mathtt{val}_{\mathcal{M},TP_{\liminf\geq 0}}(s_0) = 1$ and any FR strategy σ is such that $\mathcal{P}_{\mathcal{M},s_0,\sigma}(TP_{\liminf\geq 0}) = 0$. In particular, there are no ε-optimal k-bit Markov strategies for any $k \in \mathbb{N}$ and any $\varepsilon < 1$ for $TP_{\liminf\geq 0}$ 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 $MP_{\liminf\geq 0}$ 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 $\bot$ states. For the rest of the section, $\mathcal{M}$ is the MDP constructed in Figure 3.

▶ Remark 10. $\mathcal{M}$ 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 $\bot$ states lead to the next row.

▶ **Lemma 11.** *There exists a strategy σ such that $\mathcal{P}_{\mathcal{M},\sigma,s_0}(MP_{\liminf\geq 0}) = 1$.*

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 σ by concatenating $\sigma_{1/2}$ strategies in the sense that σ plays just like $\sigma_{1/2}$ in each gadget from each gadget's start state.

Since σ achieves at least $1/2$ in every gadget that it sees, with probability 1, runs generated by σ 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 σ mimics in every controlled state, not restarting anymore directly implies that the total payoff is eventually always ≥ 0. Hence all runs generated by σ and restarting only finitely many times satisfy $MP_{\liminf\geq 0}$. Therefore all but a nullset of runs generated by σ are winning, i.e. $\mathcal{P}_{\mathcal{M},s_0,\sigma}(MP_{\liminf\geq 0}) = 1$. ◀

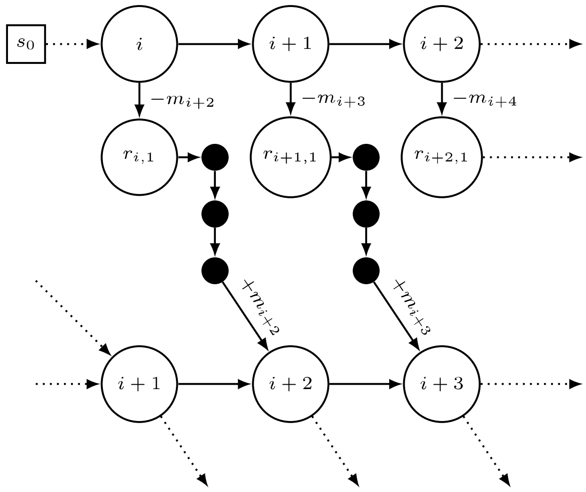

Figure 3 Each row represents a copy of the MDP depicted in Figure 2. Each white circle labeled with a number i represents the correspondingly numbered gadget (like in Figure 1) from that MDP. Now, instead of the bottom states in each gadget leading to an infinite losing chain, they lead to a restart 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 jth restart. The i indicates that the run entered the restart state from the ith 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 σ, $\mathcal{P}_{\mathcal{M},\sigma,s_0}(MP_{\liminf\geq 0}) = 0$.*

Proof sketch. Let σ be any FR strategy. We partition the runs generated by σ 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 $\mathcal{P}_{\mathcal{M},\sigma,s_0}(MP_{\liminf\geq 0}) = 0$. ◀

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

▶ **Theorem 13.** *There exists a countable, finitely branching and acyclic MDP $\mathcal{M}$ whose step counter is implicit in the state for which s_0 is almost surely winning $MP_{\liminf\geq 0}$, i.e., $\exists\hat{\sigma}\, \mathcal{P}_{\mathcal{M},s_0,\hat{\sigma}}(MP_{\liminf\geq 0}) = 1$, but every FR strategy σ is such that $\mathcal{P}_{\mathcal{M},s_0,\sigma}(MP_{\liminf\geq 0}) = 0$. 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 $TP_{\liminf\geq 0}$, giving us the following theorem.

▶ **Theorem 14.** *There exists a countable, finitely branching and acyclic MDP $\mathcal{M}$ whose step counter is implicit in the state for which s_0 is almost surely winning $TP_{\liminf\geq 0}$, i.e., $\exists\hat{\sigma}\, \mathcal{P}_{\mathcal{M},s_0,\hat{\sigma}}(TP_{\liminf\geq 0}) = 1$, but every FR strategy σ is such that $\mathcal{P}_{\mathcal{M},s_0,\sigma}(TP_{\liminf\geq 0}) = 0$. In particular, almost sure winning strategies, when they exist, cannot be chosen k-bit Markov for any $k \in \mathbb{N}$ for countable MDPs.*

4 When is a reward counter not sufficient?

In this part we show that a reward counter plus arbitrary finite memory does not suffice for (ε-)optimal strategies for $MP_{\liminf\geq 0}$, even if the MDP is finitely branching.

The same lower bound holds for $TP_{\liminf\geq 0}/PP_{\liminf\geq 0}$, but only in infinitely branching MDPs. The finitely branching case is different for $TP_{\liminf\geq 0}/PP_{\liminf\geq 0}$; 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 $\mathcal{M}_{\mathrm{RI}}$ with initial state $(s_0, 0)$ with the total reward implicit in the state such that*

- $\mathtt{val}_{\mathcal{M}_{\mathrm{RI}}, MP_{\liminf\geq 0}}((s_0, 0)) = 1$,
- *for all FR strategies σ, we have $\mathcal{P}_{\mathcal{M}_{\mathrm{RI}},(s_0,0),\sigma}(MP_{\liminf\geq 0}) = 0$.*

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

- *there exists an HD strategy σ s.t. $\mathcal{P}_{\mathcal{M}_{\mathrm{Restart}},s_0,\sigma}(MP_{\liminf\geq 0}) = 1$.*
- *for every FR strategy σ, $\mathcal{P}_{\mathcal{M}_{\mathrm{Restart}},s_0,\sigma}(MP_{\liminf\geq 0}) = 0$.*

▶ **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 σ is such that $\mathcal{P}_{\mathcal{M},s,\sigma}(TP_{\liminf\geq 0}) = 0$ and $\mathcal{P}_{\mathcal{M},s,\sigma}(PP_{\liminf\geq 0}) = 0$*
- *there exists an HD strategy σ s.t. $\mathcal{P}_{\mathcal{M},s,\sigma}(TP_{\liminf\geq 0}) = 1$ and $\mathcal{P}_{\mathcal{M},s,\sigma}(PP_{\liminf\geq 0}) = 1$.*

Hence, optimal (and even almost-surely winning) strategies and ε-optimal strategies for $TP_{\liminf\geq 0}$ and $PP_{\liminf\geq 0}$ require infinite memory beyond a reward counter.

▶ Remark 18. The MDPs from Section 3 and Section 4 show that good strategies for $MP_{\liminf\geq 0}$ 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 $MP_{\liminf\geq 0}$ 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 $\liminf$ 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 $TP_{\liminf\geq 0}$ boils down to eventually avoiding all transitions with negative reward (since negative rewards would be ≤ -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 ε-optimal MD strategies for $PP_{\liminf\geq 0}$. In turn, this yields the requisite upper bound for finitely branching $TP_{\liminf\geq 0}$, i.e., using just a reward counter.

For *infinitely branching* MDPs, a step counter suffices in order to achieve $PP_{\liminf\geq 0}$ ε-optimally. Then, by encoding the total reward into the states, this will also give us SC+RC upper bounds for $MP_{\liminf\geq 0}$ as well as infinitely branching $TP_{\liminf\geq 0}$ (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 σ' in $R(\mathcal{M})$ attaining $c \in [0,1]$ for $PP_{\liminf \geq 0}$ from $(s_0, 0)$, there exists a strategy σ attaining c for $TP_{\liminf \geq 0}$ in $\mathcal{M}$ from s_0 which uses the same memory as σ' plus a reward counter.*

▶ Remark 21. Given an MDP $\mathcal{M}$ and initial state s_0, we can construct an acyclic MDP $S(\mathcal{M})$ with initial state $(s_0, 0)$ and with the step counter implicit in the state such that MD strategies in $S(\mathcal{M})$ 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 $MP_{\liminf \geq 0}$ 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 σ' in $A(\mathcal{M})$ attaining $c \in [0,1]$ for $PP_{\liminf \geq 0}$ from $(s_0, 0, 0)$, there exists a strategy σ attaining c for $MP_{\liminf \geq 0}$ 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 ε-optimal.*

▶ **Theorem 25** ([13, Theorem 7])**.** *Let $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P, r)$ be a countable MDP, and let φ be an event that is tail in $\mathcal{M}$. Suppose for every $s \in S$ there exist ε-optimal MD strategies for φ. Then:*

1. *There exist uniform ε-optimal MD strategies for φ.*
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 $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P, r)$, `Transience` $\stackrel{\text{def}}{=} \bigwedge_{s \in S} \mathsf{FG}\neg s$.

▶ **Theorem 26** ([13, Theorem 8])**.** *In every countable MDP there exist uniform ε-optimal MD strategies for* `Transience`.

▶ **Theorem 27.** *Consider a finitely branching MDP $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P, r)$ with initial state s_0 and a $PP_{\liminf \geq 0}$ objective. Then there exist ε-optimal MD strategies.*

Proof. Let $\varepsilon > 0$. We begin by partitioning the state space into two sets, S_{safe} and $S \setminus S_{\text{safe}}$. The set S_{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 σ_{safe} for this safety objective [18, 16].

We construct a new MDP $\mathcal{M}'$ by modifying $\mathcal{M}$. We create a gadget G_{safe} composed of a sequence of new controlled states $x_0, x_1, x_2, \ldots$ where all transitions $x_i \to x_{i+1}$ have reward 0. Hence any run entering G_{safe} is winning for $PP_{\liminf\geq 0}$. We insert G_{safe} into $\mathcal{M}$ by replacing all incoming transitions to S_{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_{safe}, you can win surely by playing an optimal MD strategy for safety. So we replace S_{safe} with the surely winning gadget G_{safe}. Thus

$$\mathtt{val}_{\mathcal{M},PP_{\liminf\geq 0}}(s_0) = \mathtt{val}_{\mathcal{M}',PP_{\liminf\geq 0}}(s_0) \tag{1}$$

and if an ε-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) ε-optimal strategy σ for $PP_{\liminf\geq 0}$ from s_0 on $\mathcal{M}'$, i.e.,

$$\mathcal{P}_{\mathcal{M}',s_0,\sigma}(PP_{\liminf\geq 0}) \geq \mathtt{val}_{\mathcal{M}',PP_{\liminf\geq 0}}(s_0) - \varepsilon. \tag{2}$$

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

$$PP_{\liminf\geq 0} = \bigcap_{i\in\mathbb{N}} \mathsf{F}(\text{Safety}_i). \tag{3}$$

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

$$\mathsf{F}(\text{Safety}_i) = \bigcup_{k\in\mathbb{N}} \text{Safety}_i^k. \tag{4}$$

By continuity of measures from above we get

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

Hence for every $i \in \mathbb{N}$ and $\varepsilon_i \overset{\text{def}}{=} \varepsilon \cdot 2^{-i}$ there exists n_i such that

$$\mathcal{P}_{\mathcal{M}',s_0,\sigma}\left(\mathsf{F}(\text{Safety}_i) \cap \overline{\text{Safety}_i^{n_i}}\right) \leq \varepsilon_i. \tag{5}$$

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

▷ Claim 28.

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

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 $\text{Bubble}_n(s_0) \overset{\text{def}}{=} \{s \in S \mid d(s) \leq n\}$ 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

$$\text{Bad}_i \overset{\text{def}}{=} \{t \in \longrightarrow_{\mathcal{M}'} \mid t = s \longrightarrow_{\mathcal{M}'} s', s \notin \text{Bubble}_{n_i}(s_0) \text{ and } r(t) < -2^{-i}\}$$

be the set of transitions originating outside $\mathrm{Bubble}_{n_i}(s_0)$ whose reward is too negative. Thus a run from s_0 that satisfies $\mathrm{Safety}_i^{n_i}$ cannot use any transition in Bad_i, since (by definition of $\mathrm{Bubble}_{n_i}(s_0)$) 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 $\mathcal{M}'$ into $\mathcal{M}''$ 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 $PP_{\liminf\geq 0}$ 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 $\mathcal{M}'$ to $\mathcal{M}''$ has decreased the value of s_0 for $PP_{\liminf\geq 0}$ by at most 2ε, i.e.,

$$\mathtt{val}_{\mathcal{M}'',PP_{\liminf\geq 0}}(s_0) \geq \mathtt{val}_{\mathcal{M}',PP_{\liminf\geq 0}}(s_0) - 2\varepsilon. \tag{6}$$

Equation (6) follows from the following steps.

$$\begin{aligned}
\mathtt{val}_{\mathcal{M}'',PP_{\liminf\geq 0}}(s_0) &\geq \mathcal{P}_{\mathcal{M}'',s_0,\sigma}\left(\bigcap_{i\in\mathbb{N}} \mathrm{Safety}_i^{n_i}\right) \\
&= \mathcal{P}_{\mathcal{M}',s_0,\sigma}\left(\bigcap_{i\in\mathbb{N}} \mathrm{Safety}_i^{n_i}\right) && \text{by def. of } \mathcal{M}'' \\
&\geq \mathtt{val}_{\mathcal{M}',PP_{\liminf\geq 0}}(s_0) - 2\varepsilon && \text{by Claim 28}
\end{aligned}$$

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

▷ Claim 29.

$$\forall \sigma''.\, \mathcal{P}_{\mathcal{M}'',s_0,\sigma''}(PP_{\liminf\geq 0}) = \mathcal{P}_{\mathcal{M}'',s_0,\sigma''}(\mathtt{Transience}).$$

By Theorem 26, there exists a uniformly ε-optimal MD strategy $\widehat{\sigma}$ from s_0 for `Transience` in $\mathcal{M}''$, i.e.,

$$\mathcal{P}_{\mathcal{M}'',s_0,\hat{\sigma}}(\mathtt{Transience}) \geq \mathtt{val}_{\mathcal{M}'',\mathtt{Transience}}(s_0) - \varepsilon. \tag{7}$$

We construct an MD strategy σ^* in $\mathcal{M}$ which plays like σ_{safe} in S_{safe} and plays like $\widehat{\sigma}$ everywhere else.

$$\begin{aligned}
\mathcal{P}_{\mathcal{M},s_0,\sigma^*}(PP_{\liminf\geq 0}) &= \mathcal{P}_{\mathcal{M}',s_0,\hat{\sigma}}(PP_{\liminf\geq 0}) && \text{def. of } \sigma^* \text{ and } \sigma_{\mathrm{safe}} \\
&\geq \mathcal{P}_{\mathcal{M}'',s_0,\hat{\sigma}}(PP_{\liminf\geq 0}) && \text{new losing sink in } \mathcal{M}'' \\
&= \mathcal{P}_{\mathcal{M}'',s_0,\hat{\sigma}}(\mathtt{Transience}) && \text{by Claim 29} \\
&\geq \mathtt{val}_{\mathcal{M}'',\mathtt{Transience}}(s_0) - \varepsilon && \text{by (7)} \\
&= \mathtt{val}_{\mathcal{M}'',PP_{\liminf\geq 0}}(s_0) - \varepsilon && \text{by Claim 29} \\
&\geq \mathtt{val}_{\mathcal{M}',PP_{\liminf\geq 0}}(s_0) - 2\varepsilon - \varepsilon && \text{by (6)} \\
&= \mathtt{val}_{\mathcal{M},PP_{\liminf\geq 0}}(s_0) - 3\varepsilon && \text{by (1)}
\end{aligned}$$

Hence σ^* is a 3ε-optimal MD strategy for $PP_{\liminf\geq 0}$ from s_0 in $\mathcal{M}$ as required. ◀

▶ **Corollary 30.** *Given a finitely branching MDP $\mathcal{M}$, there exist ε-optimal strategies for $TP_{\liminf\geq 0}$ which use just a reward counter.*

Proof. By Theorem 27 and Lemma 20. ◀

▶ **Corollary 31.** *Given a finitely branching MDP $\mathcal{M}$ and initial state s_0, optimal strategies, where they exist,*

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

Proof. Since $PP_{\liminf\geq 0}$ is tail, the first claim follows from Theorem 27 and Theorem 25.

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

5.2 Infinitely Branching Case

For infinitely branching MDPs, ε-optimal strategies for $PP_{\liminf\geq 0}$ 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 $\mathcal{M}$ with initial state s_0 and a $PP_{\liminf\geq 0}$ objective. For every $\varepsilon > 0$ there exist*

- *ε-optimal MD strategies in $S(\mathcal{M})$.*
- *ε-optimal deterministic Markov strategies in $\mathcal{M}$.*

▶ **Corollary 33.** *Given an MDP $\mathcal{M}$ and initial state s_0, there exist ε-optimal strategies σ for $MP_{\liminf\geq 0}$ which use just a step counter and a reward counter.*

▶ **Corollary 34.** *Given an MDP $\mathcal{M}$ with initial state s_0,*

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

▶ **Corollary 35.** *Given an MDP $\mathcal{M}$ and initial state s_0, optimal strategies, where they exist,*

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

1 Pieter Abbeel and Andrew Y. Ng. Learning first-order Markov models for control. In *Advances in Neural Information Processing Systems 17*, pages 1–8. MIT Press, 2004. URL: http://papers.nips.cc/paper/2569-learning-first-order-markov-models-for-control.

2 Galit Ashkenazi-Golan, János Flesch, Arkadi Predtetchinski, and Eilon Solan. Reachability and safety objectives in Markov decision processes on long but finite horizons. *Journal of Optimization Theory and Applications*, 185:945–965, 2020.

3 Christel Baier and Joost-Pieter Katoen. *Principles of Model Checking*. MIT Press, 2008.

4 P. Billingsley. *Probability and Measure*. Wiley, New York, NY, 1995. Third Edition.

5 Vincent D. Blondel and John N. Tsitsiklis. A survey of computational complexity results in systems and control. *Automatica*, 36(9):1249–1274, 2000.

6 Nicole Bäuerle and Ulrich Rieder. *Markov Decision Processes with Applications to Finance*. Springer-Verlag Berlin Heidelberg, 2011.

7 K. Chatterjee, L. Doyen, and T. Henzinger. A survey of stochastic games with limsup and liminf objectives. In *Proc. of ICALP*, volume 5556 of *LNCS*. Springer, 2009.

8 Krishnendu Chatterjee and Laurent Doyen. Games and Markov decision processes with mean-payoff parity and energy parity objectives. In *Proc. of MEMICS*, volume 7119 of *LNCS*, pages 37–46. Springer, 2011.

9 Edmund M. Clarke, Thomas A. Henzinger, Helmut Veith, and Roderick Bloem, editors. *Handbook of Model Checking*. Springer, 2018. doi:10.1007/978-3-319-10575-8.

10 E.M. Clarke, O. Grumberg, and D. Peled. *Model Checking*. MIT Press, December 1999.

11 János Flesch, Arkadi Predtetchinski, and William Sudderth. Simplifying optimal strategies in limsup and liminf stochastic games. *Discrete Applied Mathematics*, 251:40–56, 2018.

12 T.P. Hill and V.C. Pestien. The existence of good Markov strategies for decision processes with general payoffs. *Stoch. Processes and Appl.*, 24:61–76, 1987.

13 S. Kiefer, R. Mayr, M. Shirmohammadi, and P. Totzke. Transience in countable MDPs. In *Proc. of CONCUR*, volume 203 of *LIPIcs*, 2021. Full version at arXiv:2012.13739.

14 Stefan Kiefer, Richard Mayr, Mahsa Shirmohammadi, and Patrick Totzke. Büchi objectives in countable MDPs. In *ICALP*, volume 132 of *LIPIcs*, pages 119:1–119:14, 2019. Full version at arXiv:1904.11573. doi:10.4230/LIPIcs.ICALP.2019.119.

15 Stefan Kiefer, Richard Mayr, Mahsa Shirmohammadi, and Patrick Totzke. Strategy Complexity of Parity Objectives in Countable MDPs. In *CONCUR*, pages 7:1–:17, 2020. doi:10.4230/LIPIcs.CONCUR.2020.7.

16 Stefan Kiefer, Richard Mayr, Mahsa Shirmohammadi, and Dominik Wojtczak. Parity Objectives in Countable MDPs. In *LICS*. IEEE, 2017. doi:10.1109/LICS.2017.8005100.

17 Richard Mayr and Eric Munday. Strategy Complexity of Mean Payoff, Total Payoff and Point Payoff Objectives in Countable MDPs. In *Proc. of CONCUR*, volume 203 of *LIPIcs*, 2021. Full version at arXiv:2107.03287.

18 Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994.

19 S. M. Ross. *Introduction to Stochastic Dynamic Programming*. Academic Press, New York, 1983.

20 Manfred Schäl. Markov decision processes in finance and dynamic options. In *Handbook of Markov Decision Processes*, pages 461–487. Springer, 2002.

21 Olivier Sigaud and Olivier Buffet. *Markov Decision Processes in Artificial Intelligence*. John Wiley & Sons, 2013.

22 William D. Sudderth. Optimal Markov strategies. *Decisions in Economics and Finance*, 43:43–54, 2020.

23 R.S. Sutton and A.G Barto. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning. MIT Press, 2018.

24 M.Y. Vardi. Automatic verification of probabilistic concurrent finite-state programs. In *Proc. of FOCS'85*, pages 327–338, 1985.