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 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

Related Version This is the full version of a CONCUR 2021 paper [17].

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 $r_0 r_1 r_2 \ldots$. 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 $r_0 r_1 r_2 \ldots$ 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 ≥ 0.

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. Appendix E.

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.

	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

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.

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 the Appendix.

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 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 $s_n \to s_{n+1}$ for all $n \in \mathbb{N}$ with reward $(-1)^n 2^n$ 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] (adapted in Figure 4) shows that there are no optimal MD (memoryless deterministic) strategies for lim inf/lim sup mean payoff. [19, Counter-example 1.3] shows that there are not even ε-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 $f : S \to [0,1]$ 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 $Runs_\mathcal{M}^{s_0}$ be the set of all runs from s_0 in the MDP $\mathcal{M}$. A *partial run* is a finite prefix of a run, $pRuns_\mathcal{M}^{s_0}$ is the set of all partial runs from s_0 and $pRuns_\mathcal{M}$ the set of partial runs from any state.

We write $\rho_s(i) \stackrel{\text{def}}{=} s_i$ for the i-th state along ρ and $\rho_e(i) \stackrel{\text{def}}{=} e_i$ 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 $\sigma : pRuns_\mathcal{M} \cdot S_\Box \to \mathcal{D}(S)$ 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* $s_0 e_0 s_1 e_1 \ldots s_n Runs_\mathcal{M}^{s_n}$: if $s_0 e_0 s_1 e_1 \ldots s_n$

is not a partial run consistent with σ then $\mathcal{P}_{\mathcal{M},s_0,\sigma}(s_0e_0s_1e_1\dots s_nRuns^{s_n}_{\mathcal{M}}) \stackrel{\text{def}}{=} 0$. Otherwise, $\mathcal{P}_{\mathcal{M},s_0,\sigma}(s_0e_0s_1e_1\dots s_nRuns^{s_n}_{\mathcal{M}}) \stackrel{\text{def}}{=} \prod_{i=0}^{n-1} \bar{\sigma}(s_0e_0s_1\dots s_i)(s_{i+1})$, 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 Appendix A 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.

▶ 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.

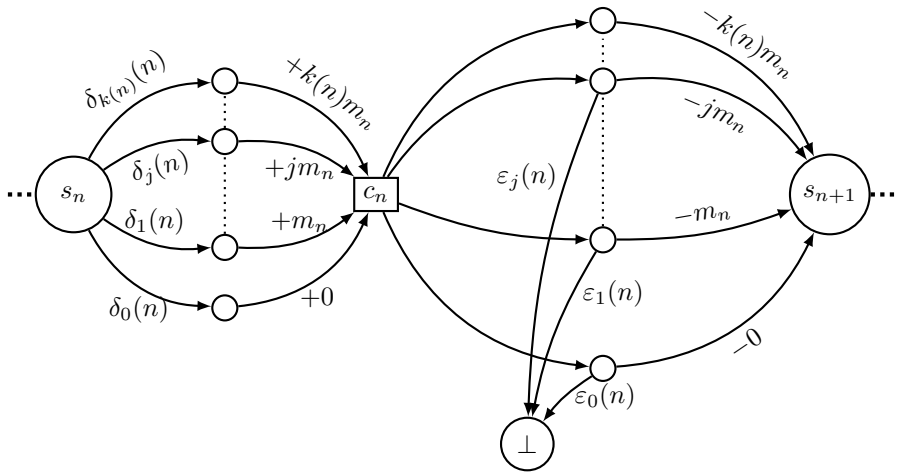

Figure 1 A typical building block with $k(n)+1$ choices, first random then controlled. The number of choices $k(n)+1$ grows unboundedly with n. This is the n-th building block of the MDP in Figure 2. The $\delta_i(n)$ and $\varepsilon_i(n)$ are probabilities depending on n and the $\pm im_n$ are transition 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 $k(n)$.

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.

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

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

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

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

▶ **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:*

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

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

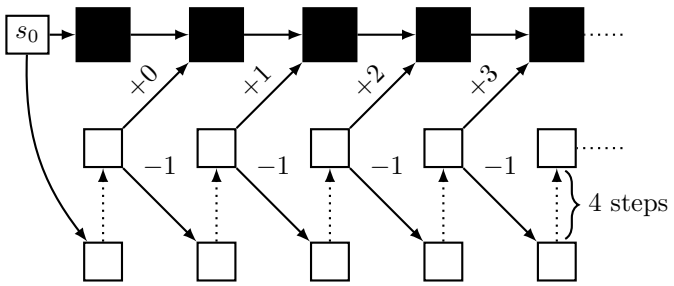

Figure 2 The buildings blocks from Figure 1 represented by black boxes are chained together (n increases as you go to the right). The chain of white boxes allows to skip arbitrarily long prefixes while preserving path length. The positive rewards from the white states to the black boxes reimburse the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

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 $k(n) \stackrel{def}{=} h^{-1}(n)$. 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 Lemma 6 and Claim 39 respectively. Notation: $N^* \stackrel{def}{=} min\{n \in \mathbb{N} : k(n) = 1\}$. *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 $m_n \stackrel{def}{=} 2k(n) \sum_{i=N^*}^{n-1} m_i$, with $m_{N^*} \stackrel{def}{=} 1$ 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 $\bot$ state. Formally, $\mathcal{P}_{\mathcal{M},s_0,\sigma_\varepsilon}(MP_{\liminf \geq 0} \wedge \mathsf{G}(\neg \bot)) = \mathcal{P}_{\mathcal{M},s_0,\sigma_\varepsilon}(\mathsf{G}(\neg \bot)) \geq 1 - \varepsilon$. So in particular, $\mathtt{val}_{\mathcal{M},MP_{\liminf \geq 0}}(s_0) = 1$.*

Proof sketch. (Full proof in Appendix B.) 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 $\bot$ 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 $\bot$ 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 $\geq 1 - \varepsilon$) by Proposition 37. 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. (Full proof in Appendix B.) 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, by Proposition 36, $\prod_{n\geq N'}(1-e_n) = 0$. 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. (Full proof in Appendix B.) 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. $\mathcal{P}_{\mathcal{M},s_0,\sigma}(MP_{\liminf\geq 0}) = 1$. ◀

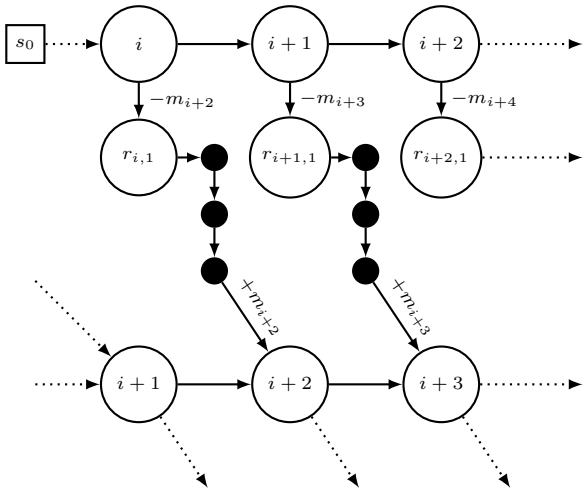

Figure 3 Each row represents a copy of the MDP depicted in Figure 2. Each white circle labeled with a number i represents the correspondingly numbered gadget (like in Figure 1) from that MDP. Now, instead of the bottom states in each gadget leading to an infinite losing chain, they lead to a restart 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. (Full proof in Appendix B.) 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 Appendix C.

▶ **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 for which $\mathtt{val}_{\mathcal{M}, MP_{\liminf \geq 0}}(s_0) = 1$ and any FR strategy σ is such that $\mathcal{P}_{\mathcal{M}_{\mathrm{Restart}}, s_0, \sigma}(MP_{\liminf \geq 0}) = 0$.*

▶ **Theorem 17.** *There exists an infinitely branching MDP $\mathcal{M}$ as in Figure 7 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 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 $TP_{\liminf \geq 0}$ 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; cf. Definition 45 for a formal definition.

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 $R(\mathcal{M})$); cf. Definition 46 for a formal definition.

▶ **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 $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P, r)$, $\texttt{Transience} \stackrel{\text{def}}{=} \bigwedge_{s \in S} \mathsf{FG}\neg s$.

▶ **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 $\text{Safety}_i^k \stackrel{\text{def}}{=} \mathsf{F}^{\leq k}(\text{Safety}_i)$ 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 $\varepsilon_i \stackrel{\text{def}}{=} \varepsilon \cdot 2^{-i}$ 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 Appendix D.1).

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

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

be the set of transitions originating outside $\text{Bubble}_{n_i}(s_0)$ whose reward is too negative. Thus a run from s_0 that satisfies $\text{Safety}_i^{n_i}$ cannot use any transition in Bad_i, since (by definition of $\text{Bubble}_{n_i}(s_0)$) they would come after the n_i-th step.

Now we create a new state $\bot$ 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 $\bot$ for every i. Notice that any run visiting $\bot$ 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.

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

In the next step (proof in Appendix D.1) 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.,

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

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

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

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 Appendix D.2.

▶ **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 $\liminf$ 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. URL: https://doi.org/10.1007/978-3-319-10575-8, doi:10.1007/978-3-319-10575-8.

10 E.M. Clarke, O. Grumberg, and D. Peled. *Model Checking*. MIT Press, Dec. 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 https://arxiv.org/abs/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 https://arxiv.org/abs/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.

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.

A Introduction to Strategy Complexity

A simple example.

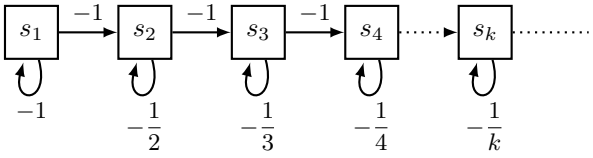

Figure 4 Adapted from [18, Example 8.10.2]. While there is no optimal MD (memoryless deterministic) strategy, the following strategy is optimal for lim inf/lim sup mean payoff: Loop $\exp(\exp(k))$ many times in state s_k for all k. In this particular example, this can be implemented with either just a step counter or just a reward counter, but in general both are needed; cf. Table 1.

Memory and strategies.

We formalize the amount of *memory* needed to implement strategies. Let M be a countable set of memory modes, and let $\tau : \mathsf{M} \times S \to \mathcal{D}(\mathsf{M} \times S)$ be a function that meets the following two conditions: for all modes $\mathsf{m} \in \mathsf{M}$,

- for all controlled states $s \in S_\Box$, the distribution $\tau(\mathsf{m}, s)$ is over $\mathsf{M} \times \{s' \mid s \longrightarrow s'\}$.
- for all random states $s \in S_\bigcirc$, and $s' \in S$, we have $\sum_{\mathsf{m}' \in \mathsf{M}} \tau(\mathsf{m}, s)(\mathsf{m}', s') = P(s)(s')$.

The function τ together with an initial memory mode m_0 induce a strategy σ_τ as follows. Consider the Markov chain with the set $\mathsf{M} \times S$ of states and the probability function τ. A sequence $\rho = s_0 \cdots s_i$ corresponds to a set $H(\rho) = \{(\mathsf{m}_0, s_0) \cdots (\mathsf{m}_i, s_i) \mid \mathsf{m}_0, \ldots, \mathsf{m}_i \in \mathsf{M}\}$ of runs in this Markov chain. Each $\rho s \in s_0 S^* S_\Box$ induces a probability distribution $\mu_{\rho s} \in \mathcal{D}(\mathsf{M})$, the probability of being in state (m, s) conditioned on having taken some partial run from $H(\rho s)$. We define σ_τ such that $\sigma_\tau(\rho s)(s') = \sum_{\mathsf{m}, \mathsf{m}' \in \mathsf{M}} \mu_{\rho s}(\mathsf{m}) \tau(\mathsf{m}, s)(\mathsf{m}', s')$ for all $\rho s \in S^* S_\Box$ and all $s' \in S$.

We say that a strategy σ can be *implemented* with memory M if there exist $\mathsf{m}_0 \in \mathsf{M}$ and τ such that $\sigma_\tau = \sigma$.

B Missing proofs from Section 3

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

Proof. Recall that Tower(i) is i repeated exponentials. Thus, log(Tower(i))=Tower($i-1$).

When checking whether probabilities in a given gadget are well defined, first we choose a gadget. The choice of gadget gives us a branching degree $k(n)+1$ which in turn lower bounds the value of n in that gadget. So for a branching degree of $k(n)+1$, we have n lower bounded by Tower$(k(n)+1)$ by definition of $k(n)$.

We need to show that $\sum_{i=0}^{k(n)-1} \delta_i(n) \leq 1$. Indeed, we have that:

$$\sum_{i=0}^{k(n)-1} \delta_i(n) \leq \sum_{i=0}^{k(n)-1} \frac{1}{\log_{i+1}(\text{Tower}(k(n)+1))} = \sum_{i=1}^{k(n)} \frac{1}{\text{Tower}(i)} < \sum_{i=1}^{k(n)} \frac{1}{e^i} < \sum_{i=1}^{k(n)} \frac{1}{2^i} < 1.$$

Hence, for $k(n) \geq 1$, the transition probabilities are well defined, i.e. $\delta_0(n), \delta_1(n), ..., \delta_{k(n)}(n)$ do indeed sum to 1. ◀

▶ **Proposition 36.** *Given an infinite sequence of real numbers a_n with $0 \leq a_n \leq 1$, we have*

$$\prod_{n=1}^{\infty}(1-a_n) > 0 \quad \Leftrightarrow \quad \sum_{n=1}^{\infty} a_n < \infty.$$

Proof. In the case where a_n does not converge to zero, the property is trivial. In the case where $a_n \to 0$, it is shown by taking the logarithm of the product and using the limit comparison test as follows.

Taking the logarithm of the product gives the series

$$\sum_{n=1}^{\infty} \ln(1-a_n)$$

whose convergence (to a finite number ≤ 0) is equivalent to the positivity of the product. It is also equivalent to the convergence (to a number ≥ 0) of its negation $\sum_{n=1}^{\infty} -\ln(1-a_n)$. But observe that (by L'Hôpital's rule)

$$\lim_{x\to 0} \frac{-\ln(1-x)}{x} = 1.$$

Since $a_n \to 0$ we have

$$\lim_{n\to\infty} \frac{-\ln(1-a_n)}{a_n} = 1.$$

By the limit comparison test, the series $\sum_{n=1}^{\infty} -\ln(1-a_n)$ converges if and only if the series $\sum_{n=1}^{\infty} a_n$ converges. ◀

▶ **Proposition 37.** *Given an infinite sequence of real numbers a_n with $0 \leq a_n \leq 1$,*

$$\prod_{n=1}^{\infty} a_n > 0 \quad \Rightarrow \quad \forall \varepsilon > 0\, \exists N.\ \prod_{n=N}^{\infty} a_n \geq (1-\varepsilon).$$

Proof. Since $\prod_{n=1}^{\infty} a_n > 0$, by taking the logarithm we obtain $\sum_{n=1}^{\infty} \ln(a_n) > -\infty$. Thus for every $\delta > 0$ there exists an N s.t. $\sum_{n=N}^{\infty} \ln(a_n) \geq -\delta$. By exponentiation we obtain $\prod_{n=N}^{\infty} a_n \geq \exp(-\delta)$. By picking $\delta = -\ln(1-\varepsilon)$ the result follows. ◀

▶ **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 $\bot$ state. Formally, $\mathcal{P}_{\mathcal{M},s_0,\sigma_\varepsilon}(MP_{\liminf \geq 0} \wedge \mathsf{G}(\neg\ \bot)) = \mathcal{P}_{\mathcal{M},s_0,\sigma_\varepsilon}(\mathsf{G}(\neg\ \bot)) \geq 1-\varepsilon$. So in particular, $\mathtt{val}_{\mathcal{M},MP_{\liminf \geq 0}}(s_0) = 1$.*

Proof. We define a strategy σ which in c_n always mimics the choice in s_n. We first prove that playing this way gives us a positive chance of winning. Then we show that there are strategies σ_ε that attain $1-\varepsilon$ from s_0 without hitting a $\bot$ state. This implies in particular that $\mathtt{val}_{\mathcal{M},MP_{\liminf \geq 0}}(s_0) = 1$.

Playing according to σ, the only way to lose is by dropping into the $\bot$ state. This is because by mimicking, the player finishes each gadget with a reward of 0. In the n-th gadget, the chance of reaching the $\bot$ state is $\sum_{j=0}^{k(n)-1} \delta_j(n) \cdot \varepsilon_j(n)$. Thus, the probability of surviving while playing in all the gadgets is

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

However, by Proposition 36, this product is strictly greater than 0 if and only if the sum

$$\sum_{n \geq N^*} \left(\sum_{i=0}^{k(n)-1} \delta_i(n)\varepsilon_i(n)\right)$$

is finite. With some rearranging exploiting the definition of $k(n)$ we see that this is indeed the case:

$$\begin{aligned}
&\sum_{n \geq N^*} \left(\sum_{i=0}^{k(n)-1} \delta_i(n)\varepsilon_i(n)\right) \\
\leq &\sum_{i \geq 1} \left(\sum_{n=g(i)}^{\infty} \delta_{i-1}(n)\varepsilon_{i-1}(n)\right) && \text{by definition of } k(n) \\
\leq &\sum_{i \geq 1} 2^{-i} && \text{by definition of } g(n) \\
\leq &1
\end{aligned}$$

Hence the player has a non zero chance of winning.

When playing with the ability to skip gadgets, as illustrated in Figure 2, all runs not visiting a $\bot$ state are winning since the total reward never dips below 0. Hence $\mathcal{P}_{\mathcal{M},s_0,\sigma_\varepsilon}(MP_{\liminf \geq 0} \wedge \neg\ \bot) = \mathcal{P}_{\mathcal{M},s_0,\sigma_\varepsilon}(\neg\ \bot)$. Thus the idea is to skip an arbitrarily long prefix of gadgets to push the chance of winning ε close to 1 by pushing the chance of visiting a $\bot$ state ε close to 0. From the N-th state, for $N \geq N^*$, the chance of winning is

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

By Proposition 37 this can be made arbitrarily close to 1 by choosing N sufficiently large.

Let $N_\varepsilon \stackrel{\text{def}}{=} \min\left\{N \in \mathbb{N} \mid \prod_{n \geq N} \left(1 - \sum_{j=0}^{k(n)-1} \delta_j(n) \cdot \varepsilon_j(n)\right) \geq 1-\varepsilon\right\}$. Now define the strategy σ_ε to be the strategy that plays like σ after skipping forwards by N_ε gadgets. Thus, by definition σ_ε attains $1-\varepsilon$ for all $\varepsilon > 0$.

Thus, by playing σ_ε for an arbitrarily small ε the chance of winning must be arbitrarily close to 1. Hence, $\mathtt{val}_{\mathcal{M},MP_{\liminf \geq 0}}(s_0) = 1$. ◀

▶ **Lemma 38.** *For any sequence $\{\alpha_n\}$, where $\alpha_n \in [0,1]$ for all n, and any functions $i(n), j(n) : \mathbb{N} \to \mathbb{N}$ with $i(n), j(n) \in \{0, 1, ..., k(n) - 1\}, i(n) < j(n)$ for all n, the following sum diverges:*

$$\sum_{n=k^{-1}(2)}^{\infty} \Big(\delta_{j(n)}(n)(\alpha_n \varepsilon_{j(n)}(n) + (1-\alpha_n)\varepsilon_{i(n)}(n)) + \delta_{i(n)}(n)(\alpha_n + (1-\alpha_n)\varepsilon_{i(n)}(n))\Big). \quad (8)$$

Proof. We can narrow our focus by noticing that

$$\begin{aligned}
&\sum_{n=k^{-1}(2)}^{\infty} \Big(\delta_{j(n)}(n)(\alpha_n \varepsilon_{j(n)}(n) + (1-\alpha_n)\varepsilon_{i(n)}(n)) + \delta_{i(n)}(n)(\alpha_n + (1-\alpha_n)\varepsilon_{i(n)}(n))\Big) \\
&= \sum_{n=k^{-1}(2)}^{\infty} \alpha_n \delta_{j(n)}(n)\varepsilon_{j(n)}(n) + (1-\alpha_n)\delta_{i(n)}\varepsilon_{i(n)}(n) \qquad \text{Convergent by def. of } \delta_i(n), \varepsilon_i(n) \\
&+ \sum_{n=k^{-1}(2)}^{\infty} (1-\alpha_n)\delta_{j(n)}\varepsilon_{i(n)}(n) + \alpha_n \delta_{i(n)}(n)
\end{aligned}$$

Hence the divergence of (8) depends only on the divergence of $\sum_{n=k^{-1}(2)}^{\infty}(1-\alpha_n)\delta_{j(n)}\varepsilon_{i(n)}(n) + \alpha_n\delta_{i(n)}(n)$. No matter how the sequence $\{\alpha_n\}$ behaves, for every n we have that either $\alpha_n \geq 1/2$ or $1 - \alpha_n \geq 1/2$. Hence for every n it is the case that

$$\begin{aligned}
(1-\alpha_n)\delta_{j(n)}(n)\varepsilon_{i(n)}(n) + \alpha_n\delta_{i(n)}(n) &\geq \frac{1}{2}\delta_{j(n)}(n)\varepsilon_{i(n)}(n) \\
&\text{or} \\
&\geq \frac{1}{2}\delta_{i(n)}(n)
\end{aligned}$$

Define the function f as follows:

$$f(n) = \begin{cases} \frac{1}{2}\delta_{i(n)}(n) \text{ if } \alpha_n \geq 1/2 \\ \\ \frac{1}{2}\delta_{j(n)}(n)\varepsilon_{i(n)}(n) \text{ otherwise} \end{cases}$$

Hence no matter how $\{\alpha_n\}$ behaves, we have that

$$\sum_{n=k^{-1}(2)}^{\infty} \Big(\delta_{j(n)}(n)(\alpha_n\varepsilon_{j(n)}(n)+(1-\alpha_n)\varepsilon_{i(n)}(n))+\delta_{i(n)}(n)(\alpha_n+(1-\alpha_n)\varepsilon_{i(n)}(n))\Big) \geq \sum_{n=k^{-1}(2)}^{\infty} f(n).$$

We know that both $\sum_{n=k^{-1}(2)}^{\infty} \frac{1}{2}\delta_{j(n)}(n)\varepsilon_{i(n)}(n)$ and $\sum_{n=k^{-1}(2)}^{\infty} \frac{1}{2}\delta_{i(n)}(n)$ diverge for all $i(n), j(n) \in \{0, 1, ..., k(n) - 1\}$, $i(n) < j(n)$, as shown in Claim 39.

Thus $\sum_{n=k^{-1}(2)}^{\infty} \frac{1}{2}\delta_{j(n)}(n)\varepsilon_{i(n)}(n)$ and $\sum_{n=k^{-1}(2)}^{\infty} \frac{1}{2}\delta_{i(n)}(n)$ must also diverge no matter how $i(n)$ and $j(n)$ behave. As a result it must be the case that $\sum_{n=k^{-1}(2)}^{\infty} f(n)$ diverges. Hence (8) must be divergent as desired as $i(n)$ and $j(n)$ vary for $n \geq k^{-1}(2)$. ◀

▷ Claim 39. The sum $\sum_{n=k^{-1}(2)}^{\infty} \frac{1}{2}\delta_{j(n)}(n)\varepsilon_{i(n)}(n)$ diverges for all $i(n), j(n) \in \{0, 1, ..., k(n) - 1\}$ with $i(n) < j(n)$.

Proof. This result is not immediate because the range of values the indexing functions $i(n)$ and $j(n)$ can take grows with $k(n)$ as n increases.

Under the assumption that $i(n) < j(n)$ we have that $\delta_{j(n)}(n)\varepsilon_{i(n)}(n) \geq \delta_{j(n)}(n)\varepsilon_{j(n)-1}(n) \geq \delta_{k(n)-1}(n)\varepsilon_{k(n)-2}(n) = \varepsilon_{k(n)-1}(n)$. Thus it suffices to show that $\sum_{n=k^{-1}(2)}^{\infty} \varepsilon_{k(n)-1}(n)$ diverges:

$$\begin{aligned}
\sum_{n=k^{-1}(2)}^{\infty} \varepsilon_{k(n)-1}(n) &= \sum_{a=2}^{\infty} \sum_{n=k^{-1}(a)}^{k^{-1}(a+1)-1} \varepsilon_{a-1}(n) && \text{splitting the sum up} \\
&= \sum_{a=2}^{\infty} \sum_{n=h(a)}^{h(a+1)-1} \varepsilon_{a-1}(n) && k(n) = h^{-1}(n) \\
&\geq \sum_{a=2}^{\infty} 1 && \text{definition of } h(n)
\end{aligned}$$

Note that the definition of $h(i)$ says exactly that a block of the form $\sum_{n=h(a)}^{h(a+1)-1} \varepsilon_{a-1}(n)$ is at least 1. Hence $\sum_{n=k^{-1}(2)}^{\infty} \frac{1}{2}\delta_{j(n)}(n)\varepsilon_{i(n)}(n)$ diverges as required.

◀

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

Proof. Let σ be some FR strategy with k memory modes. Our MDP consists of a linear sequence of gadgets (Figure 1) and is in particular acyclic. The n-th gadget is entered at state s_n and takes 4 steps. Locally in the n-th gadget there are 3 possible scenarios:

(1) The random transition picks some branch i at s_n and the strategy then picks a branch $j > i$ at c_n.
By the definition of the payoffs (multiples of m_n; cf. Definition 4), this means that we see a mean payoff ≤ -1, regardless of events in past gadgets. This is because the numbers m_n grow so quickly with n that even the combined maximal possible rewards of all past gadgets are so small in comparison that they do not matter for the outcome in the n-th gadget, i.e., rewards from past gadgets cannot help to avoid seeing a mean payoff ≤ -1 in the above scenario.

(2) We reach the losing sink $\perp$ (and thus will keep seeing a mean payoff ≤ -1 forever). This happens with probability $\varepsilon_j(n)$ if the strategy picks some branch j at c_n, regardless of past events.

(3) All other cases.

As explained above, due to the definition of the rewards (Definition 4), events in past gadgets do not make the difference between (1),(2),(3) in the current gadget. It just depends on the choices of the strategy σ in the current gadget.

Let Bad_n be the event of seeing either of the two unfavorable outcomes (1) or (2) in the n-th gadget. Let p_n be the probability of Bad_n under strategy σ. Since σ has memory, the probabilities p_n are not necessarily independent. However, we show *lower bounds* $e_n \leq p_n$ that hold universally for every FR strategy σ with $\leq k$ memory modes and every n such that $k(n) > k + 1$. The lower bound e_n will hold regardless of the memory mode of σ upon entering the n-th gadget.

Memory updates. First we show that σ randomizing its memory update after observing the random transition from state s_n does *not* help to reduce the probability of event Bad_n. I.e., we show that without restriction σ can update its memory deterministically after observing the transition from state s_n.

Once in the controlled state c_n, the strategy σ can base its choice only on the current state (always c_n in the n-th gadget) and on the current memory mode. Thus, in state c_n, in each memory mode m, the strategy has to pick a distribution $\mathcal{D}^{c_n}_{\mathsf{m}}$ over the available transitions from c_n. By the finiteness of the number of memory modes of σ (just $\leq k$ by our assumption above), for each possible reward level x (obtained in the step from the preceding random transition) there is a best memory mode $\mathsf{m}(x)$ such that $\mathcal{D}^{c_n}_{\mathsf{m}(x)}$ is optimal (in the sense of minimizing the probability of event Bad_n) for that particular reward level x. (In case of a tie, just use an arbitrary tie break, e.g., some pre-defined linear order on the memory modes.)

Therefore, upon witnessing a reward level x in the random transition from state s_n, the strategy σ can minimize the probability of event Bad_n by *deterministically* setting its memory to $\mathsf{m}(x)$. Thus randomizing its memory update does not help to reduce the probability of Bad_n, and we may assume without restriction that σ updates its memory deterministically.

(Note that the above argument only works because it is local to the current gadget where we have a finite number of decisions (here just one), we have a finite number of memory modes, and a one-dimensional criterion for local optimality (minimizing the probability of event Bad_n). We do *not* claim that randomized memory updates are useless for every strategy in every MDP and every objective.)

The lower bounds e_n. Now we consider an FR strategy σ that without restriction updates its memory *deterministically* after each random choice (from state s_n) in the n-th gadget. It can still randomize its actions, however.

Let N' be the minimal number such that for all $n \geq N'$ we have $k(n) > k+1$. In particular, this implies $N' \geq k^{-1}(2)$, and thus we can apply Lemma 38 later.

Once $n \geq N'$, then by the Pigeonhole Principle there will always be a memory mode confusing at least two different transitions $i(n), j(n) \neq k(n)$ from state s_n to c_n. Note that this holds regardless of the memory mode of σ upon entering the n-th gadget. (The strategy might confuse many other scenarios, but just one confused pair $i(n), j(n) \neq k(n)$ is enough for our lower bound.) Without loss of generality, let $j(n)$ be larger of the two confused transitions, i.e., $i(n) < j(n)$. Let $i(n)$ and $j(n)$ be two functions taking values in $\{0, 1, ..., k(n)-1\}$ where $i(n) < j(n)$ for all n.

Confusing two transitions $i(n)$ and $j(n)$ from s_n to c_n (where without restriction $i(n) < j(n)$), the strategy is in the same memory mode afterwards. However, it can still randomize its choices in state c_n. To prove our lower bound on the probability of Bad_n, it suffices to consider the case where the strategy only randomizes over the outgoing transitions $i(n)$ and $j(n)$ from state c_n. This is because, by Claim 40, every other behavior would perform even worse, in the sense of yielding a higher probability of Bad_n.

That is to say that the strategy picks the higher $j(n)$-th branch with some probability α_n and the lower $i(n)$-th branch with probability $1-\alpha_n$. (We leave the probabilities α_n unspecified here. Using Lemma 38, we'll show that our result holds regardless of their values.)

The local chance of the event Bad_n is then lower bounded by

$$e_n \stackrel{\text{def}}{=} \delta_{j(n)}(n)(\alpha_n \varepsilon_{j(n)}(n) + (1-\alpha_n)\varepsilon_{i(n)}(n)) + \delta_{i(n)}(n)(\alpha_n + (1-\alpha_n)\varepsilon_{i(n)}(n)).$$

The term above just expresses a case distinction. In the first scenario, the random transition chooses the $j(n)$-th branch (with probability $\delta_{j(n)}(n)$) and then the strategy

chooses the $j(n)$-th branch with probability α_n and the lower $i(n)$-th branch with probability $1-\alpha_n$, and you obtain the respective chances of reaching the sink $\bot$. In the second scenario, the random transition chooses the $i(n)$-th branch (with probability $\delta_{i(n)}(n)$). If the strategy then chooses the higher $j(n)$-th branch (with probability α_n) then we have outcome (1), yielding a mean payoff ≤ -1. If the strategy chooses the $i(n)$-th branch (with probability $1-\alpha_n$) then we still have a chance of $\varepsilon_{i(n)}(n)$ of reaching the sink.

Since, as shown above, randomized memory updates do not help to reduce the probability of Bad_n, the lower bound e_n for deterministic updates carries over to the general case. Thus, even for general randomized FR strategies σ with k memory modes, the probability of event Bad_n in the n-th gadget (for $n \geq N'$) is lower bounded by e_n, regardless of the memory mode m upon entering the gadget and regardless of events in past gadgets. We write $\sigma[\mathsf{m}]$ for the strategy σ in memory mode m and obtain

$$\forall n \geq N'.\ \forall \mathsf{m}.\ \mathcal{P}_{\mathcal{M},\sigma[m],s_n}(Bad_n) \geq e_n \tag{9}$$

The final step. Let $Bad \stackrel{\text{def}}{=} \cup_n Bad_n$.

Since $i(n), j(n) \neq k(n)$ and $N' \geq k^{-1}(2)$, we apply Lemma 38 to conclude that the series $\sum_{n=N'}^{\infty} e_n = \sum_{n=N'}^{\infty} \delta_{j(n)}(n)(\alpha_n \varepsilon_{j(n)}(n) + (1-\alpha_n)\varepsilon_{i(n)}(n)) + \delta_{i(n)}(n)(\alpha_n + (1-\alpha_n)\varepsilon_{i(n)}(n))$ is divergent, regardless of the behavior of $i(n), j(n)$ or the sequence $\{\alpha_n\}$.

Finally, we obtain

$$\begin{aligned}
&\mathcal{P}_{\mathcal{M},\sigma,s_0}(MP_{\liminf \geq 0}) \\
&\leq \mathcal{P}_{\mathcal{M},\sigma,s_0}(\mathsf{FG}\neg Bad) && \text{set inclusion} \\
&= \mathcal{P}_{\mathcal{M},\sigma,s_0}\left(\bigcup_l \mathsf{F}^{\leq l}\mathsf{G}\neg Bad\right) && \text{def. of } \mathsf{F} \\
&= \lim_{l\to\infty} \mathcal{P}_{\mathcal{M},\sigma,s_0}(\mathsf{F}^{\leq l}\mathsf{G}\neg Bad) && \text{continuity of measures} \\
&\leq \lim_{l\to\infty} \mathcal{P}_{\mathcal{M},\sigma,s_0}\left(\bigcap_{n\geq l/4} \neg Bad_n\right) && \text{4 steps per gadget} \\
&\leq \lim_{4N'\leq l\to\infty} \prod_{n\geq l/4\geq N'} (\max_{\mathsf{m}} \mathcal{P}_{\mathcal{M},\sigma[\mathsf{m}],s_n}(\neg Bad_n)) && \text{linear sequence of gadgets, finite memory, and past events do not help to avoid } Bad_n \\
&\leq \lim_{4N'\leq l\to\infty} \prod_{n\geq l/4\geq N'} (1-e_n) && \text{by (9)} \\
&= \lim_{4N'\leq l\to\infty} 0 && \text{divergence of } \sum_{n=N'}^{\infty} e_n \text{ and Proposition 36} \\
&= 0
\end{aligned}$$

◀

▷ Claim 40. Assume that the transitions $i(n)$ and $j(n)$ (with $i(n) < j(n)$) leading to state c_n are confused in the memory of the strategy. Then we can assume without restriction that the strategy only plays transitions $i(n)$ and $j(n)$ with nonzero probability from state c_n, since every other behavior yields a higher probability of the event Bad_n (cf. Figure 5).

Proof. When confusing transitions $i(n)$ and $j(n)$ with $i(n) < j(n)$, the player's choice of transition from c_n can be broken down into 5 distinct cases. The player can choose transition $x(n)$ as follows.

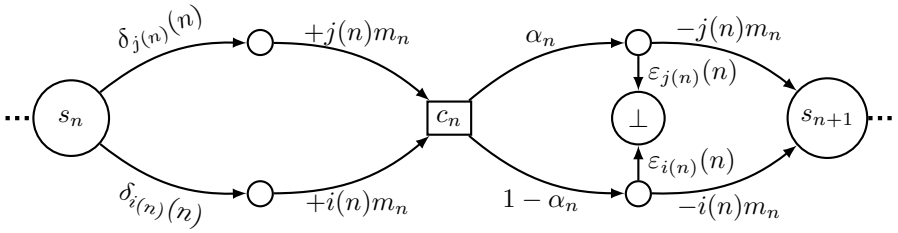

Figure 5 When transitions $i(n)$ and $j(n)$ are confused in the player's memory, the player's choice is at least as bad as the reduced play in this simplified gadget.

1. $x(n) = i(n)$
2. $x(n) = j(n)$
3. $x(n) > j(n)$
4. $x(n) < i(n)$
5. $i(n) < x(n) < j(n)$

Case 1 leads to a probability of Bad_n of $\delta_{j(n)}(n)\varepsilon_{i(n)}(n) + \delta_{i(n)}(n)\varepsilon_{i(n)}(n)$.

Case 2 leads to a probability of Bad_n of $\delta_{j(n)}(n)\varepsilon_{j(n)}(n) + \delta_{i(n)}(n)$.

Case 3 leads to a mean payoff ≤ -1 (and thus Bad_n) with probability 1. This is the worst possible case.

Case 4 leads to a probability of Bad_n of $\delta_{j(n)}(n)\varepsilon_{x(n)}(n)+\delta_{i(n)}(n)\varepsilon_{x(n)}(n) > \delta_{j(n)}(n)\varepsilon_{i(n)}(n)+\delta_{i(n)}(n)\varepsilon_{i(n)}(n)$, i.e., this is worse than Case 1.

Case 5 leads to a probability of Bad_n of $\delta_{j(n)}(n)\varepsilon_{x(n)}(n) + \delta_{i(n)}(n) > \delta_{j(n)}(n)\varepsilon_{j(n)}(n) + \delta_{i(n)}(n)$, i.e., this is worse than Case 2.

Hence, without restriction we can assume that only cases 1 and 2 will get played with positive probability, that is to say that in state c_n the strategy will only randomize over the outgoing transitions $i(n)$ and $j(n)$. ◀

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

Proof. We will show that there exists a strategy σ that satisfies the mean payoff objective with probability 1 from s_0. Towards this objective we recall the strategy $\sigma_{1/2}$ defined in Lemma 6. In a given gadget of this MDP with restarts, playing $\sigma_{1/2}$ in said gadget, there is a probability of at most $1/2$ of restarting in that gadget. We then construct 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.

Let $\mathfrak{R}$ be the set of runs induced by σ from s_0. We partition $\mathfrak{R}$ into the sets $\mathfrak{R}_i$ and $\mathfrak{R}_\infty$ of runs such that $\mathfrak{R} = (\bigcup_{i=0}^{\infty} \mathfrak{R}_i) \cup \mathfrak{R}_\infty$. We define for $i = 0$

$$\mathfrak{R}_0 \stackrel{\text{def}}{=} \{\rho \in \mathfrak{R} \mid \forall \ell \in \mathbb{N}.\, \neg \mathsf{F}(r_{\ell,1})\},$$

for $i \geq 1$

$$\mathfrak{R}_i \stackrel{\text{def}}{=} \{\rho \in \mathfrak{R} \mid \exists j \in \mathbb{N}.\, \mathsf{F}(r_{j,i}) \wedge \forall \ell \in \mathbb{N}.\, \neg \mathsf{F}(r_{\ell,i+1})\}$$

and

$$\mathfrak{R}_\infty \stackrel{\text{def}}{=} \{\rho \in \mathfrak{R} \mid \forall i \in \mathbb{N}\ \exists j \in \mathbb{N}.\, \mathsf{F}(r_{j,i})\}.$$

That is to say for all $i \in \mathbb{N}$, $\mathfrak{R}_i$ is the set of runs in $\mathfrak{R}$ that restart exactly i times and $\mathfrak{R}_\infty$ is the set of runs in $\mathfrak{R}$ that restart infinitely many times.

We go on to define the sets of runs $\mathfrak{R}_{\geq i} \stackrel{\text{def}}{=} \bigcup_{j=i}^{\infty} \mathfrak{R}_j$ which are those runs which restart at least i times. In particular note that $\mathfrak{R}_\infty = \bigcap_{i=0}^{\infty} \mathfrak{R}_{\geq i}$ and $\mathfrak{R}_{\geq i+1} \subseteq \mathfrak{R}_{\geq i}$.

By construction, any run $\rho \in \mathfrak{R}_\infty$ is losing since the negative reward that is collected upon restarting instantly brings the mean payoff below -1 by definition of m_n. Thus restarting infinitely many times translates directly into the mean payoff dropping below -1 infinitely many times and thus a strictly negative lim inf mean payoff. As a result it must be the case that $\mathfrak{R}_\infty \subseteq \neg MP_{\liminf \geq 0}$.

After every restart, the negative reward is reimbursed. Intuitively, going through finitely many restarts does not damage the chances of winning. We now show that, except for a nullset, the runs restarting only finitely many times satisfy the objective. Indeed, every run with only finitely many restarts must spend an infinite tail in some final gadget in which it does not restart. In this final gadget, the strategy plays just like $\sigma_{1/2}$, which means that it mimics the random choice in every controlled state. Since, by assumption, there are no more restarts, we obtain $\mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_i) = \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_i \wedge \forall j \in \mathbb{N}, \mathsf{G}(\neg r_{j,i+1}))$. We then apply Lemma 6 to obtain that

$$\mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_i) = \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_i \wedge \forall j \in \mathbb{N}, \mathsf{G}(\neg r_{j,i+1})) = \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_i \wedge MP_{\liminf \geq 0}). \tag{10}$$

In other words, except for a nullset, the run restarting finitely often (here i times) satisfy $MP_{\liminf \geq 0}$. Furthermore, notice that from this observation, the sets $\mathfrak{R}_i$ partition the set of winning runs.

We show now that $\mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_\infty) = 0$. We do so firstly by showing by induction that $\mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_{\geq i}) \leq 2^{-i}$ for $i \geq 1$, then applying the continuity of measures from above to obtain that $\mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_\infty) = 0$.

Our base case is $i = 1$. $\mathfrak{R}$, by definition of σ, is the set of runs induced by playing $\sigma_{1/2}$ in every gadget. By Lemma 6 σ attains $\geq 1/2$ in every gadget. Therefore in particular the probability of a run leaving the first gadget is no more than $1/2$, i.e. $\mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_{\geq 1}) \leq 1/2$.

Now suppose that $\mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_{\geq i}) \leq 2^{-i}$. After restarting at least i times, the probability of a run restarting at least once more is still $\leq 1/2$ since the strategy being played in every gadget is $\sigma_{1/2}$. Hence

$$\mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_{\geq i+1}) \leq \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_{\geq i}) \cdot \frac{1}{2} \leq 2^{-(i+1)}$$

which is what we wanted.

Now we use the fact that $\mathfrak{R}_\infty = \bigcap_{i=0}^{\infty} \mathfrak{R}_{\geq i}$ and $\mathfrak{R}_{\geq i+1} \subseteq \mathfrak{R}_{\geq i}$ to apply continuity of measures from above and obtain:

$$\mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_\infty) = \mathcal{P}_{\mathcal{M},s_0,\sigma}\left(\bigcap_{i=0}^{\infty} \mathfrak{R}_{\geq i}\right) = \lim_{i \to \infty} \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_{\geq i}) \leq \lim_{i \to \infty} 2^{-i} = 0.$$

Hence $\mathfrak{R}_\infty$ is a null set.

We can now write down the following:

$$
\begin{aligned}
1 &= \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}) \\
&= \left(\sum_{i=0}^{\infty} \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_i)\right) + \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_\infty) && \text{by partition of } \mathfrak{R} \\
&= \left(\sum_{i=0}^{\infty} \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_i \wedge MP_{\liminf \geq 0})\right) + \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_\infty) && \text{by Equation (10)} \\
&= \left(\sum_{i=0}^{\infty} \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_i \wedge MP_{\liminf \geq 0})\right) + \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_\infty \wedge MP_{\liminf \geq 0}) && \text{by } \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_\infty) = 0 \\
&= \mathcal{P}_{\mathcal{M},s_0,\sigma}(MP_{\liminf \geq 0}) && \text{by partition of } MP_{\liminf \geq 0}
\end{aligned}
$$

Thus $\mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}) = \mathcal{P}_{\mathcal{M},s_0,\sigma}(MP_{\liminf \geq 0}) = 1$, i.e. σ wins almost surely. ◀

▶ **Lemma 12.** *For any FR strategy σ, $\mathcal{P}_{\mathcal{M},\sigma,s_0}(MP_{\liminf \geq 0}) = 0$.*

Proof. There are two ways to lose when playing in this MDP: either the mean payoff dips below -1 infinitely often because the run takes infinitely many restarts, or the run only takes finitely many restarts, but the mean payoff drops below -1 infinitely many times in the last copy of the gadget that the run stays in. Recall that in Lemma 7 we showed that any FR strategy with probability 1 either restarts or lets the mean payoff dip below -1 infinitely often.

Let σ be any FR strategy and let $\mathfrak{R}$ to be the set of runs induced by σ from s_0. We partition $\mathfrak{R}$ into the sets $\mathfrak{R}_i$ and $\mathfrak{R}_\infty$ of runs such that $\mathfrak{R} = (\bigcup_{i=0}^{\infty} \mathfrak{R}_i) \cup \mathfrak{R}_\infty$. Where we define for $i = 0$

$$\mathfrak{R}_0 \stackrel{\text{def}}{=} \{\rho \in \mathfrak{R} \mid \forall \ell \in \mathbb{N}, \neg \mathsf{F}(r_{\ell,1})\},$$

for $i \geq 1$

$$\mathfrak{R}_i \stackrel{\text{def}}{=} \{\rho \in \mathfrak{R} \mid \exists j \in \mathbb{N}, \mathsf{F}(r_{j,i}) \wedge \forall \ell \in \mathbb{N}, \neg \mathsf{F}(r_{\ell,i+1})\}$$

and

$$\mathfrak{R}_\infty \stackrel{\text{def}}{=} \{\rho \in \mathfrak{R} \mid \forall i, \exists j\ \mathrm{F}(r_{j,i})\}.$$

That is to say for all $i \in \mathbb{N}$, $\mathfrak{R}_i$ is the set of runs in $\mathfrak{R}$ that restart exactly i times and $\mathfrak{R}_\infty$ is the set of runs in $\mathfrak{R}$ that restart infinitely many times.

We go on to define the sets of runs $\mathfrak{R}_{\geq i} \stackrel{\text{def}}{=} \bigcup_{j=i}^{\infty} \mathfrak{R}_j$ which are those runs which restart at least i times. In particular note that $\mathfrak{R}_\infty = \bigcap_{i=0}^{\infty} \mathfrak{R}_{\geq i}$ and $\mathfrak{R}_{\geq i+1} \subseteq \mathfrak{R}_{\geq i}$.

Note that any run in $\mathfrak{R}_\infty$ is losing by construction. The negative reward that is collected upon restarting instantly brings the mean payoff below -1 by definition of m_n. Thus restarting infinitely many times translates directly into the mean payoff dropping below -1 infinitely many times. Thus $\mathfrak{R}_\infty \subseteq \neg MP_{\liminf \geq 0}$ and so it follows that $\mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_\infty) = \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_\infty \wedge \neg MP_{\liminf \geq 0})$. Since the sets $\mathfrak{R}_i$ and $\mathfrak{R}_\infty$ partition $\mathfrak{R}$ we have that:

$$\mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}) = \left(\sum_{i=0}^{\infty} \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_i)\right) + \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_\infty).$$

It remains to show that every set $\mathfrak{R}_i$ is almost surely losing, i.e. $\mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_i) = \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_i \wedge \neg MP_{\liminf \geq 0})$. Consider a run $\rho \in \mathfrak{R}_i$. By definition it restarts exactly i times. As a result, it spends infinitely long in the $i+1$st gadget. Because σ is an FR strategy, it must be the

case that any substrategy σ^* induced by σ that is played in a given gadget is also an FR strategy. This allows us to apply Lemma 7 to obtain that

$$\mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_i) = \mathcal{P}_{\mathcal{M},s_0,\sigma}\left(\mathfrak{R}_i \wedge (\neg MP_{\liminf\geq 0} \vee \exists j \in \mathbb{N}, \mathsf{F}(r_{j,i+1}))\right). \tag{11}$$

However, any run $\rho \in \mathfrak{R}_i$ never sees any state $r_{j,i+1}$ for any j by definition. Therefore it follows that

$$\mathcal{P}_{\mathcal{M},s_0,\sigma}\left(\mathfrak{R}_i \wedge (\neg MP_{\liminf\geq 0} \vee \exists j \in \mathbb{N}, \mathsf{F}(r_{j,i+1}))\right) = \mathcal{P}_{\mathcal{M},s_0,\sigma}\left(\mathfrak{R}_i \wedge (\neg MP_{\liminf\geq 0})\right)$$

Hence $\mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_i) = \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_i \wedge \neg MP_{\liminf\geq 0})$ as required.

As a result we have that

$$\begin{aligned}
1 &= \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}) \\
&= \left(\sum_{i=0}^{\infty} \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_i)\right) + \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_\infty) && \text{by partition of } \mathfrak{R} \\
&= \left(\sum_{i=0}^{\infty} \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_i \wedge \neg MP_{\liminf\geq 0})\right) + \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R}_\infty \wedge \neg MP_{\liminf\geq 0}) && \text{by Equation (11)} \\
&= \mathcal{P}_{\mathcal{M},s_0,\sigma}(\neg MP_{\liminf\geq 0}) && \text{by partition of } \mathfrak{R}
\end{aligned}$$

That is to say that for any FR strategy σ, $\mathcal{P}_{\mathcal{M},s_0,\sigma}(MP_{\liminf\geq 0}) = 0$. ◀

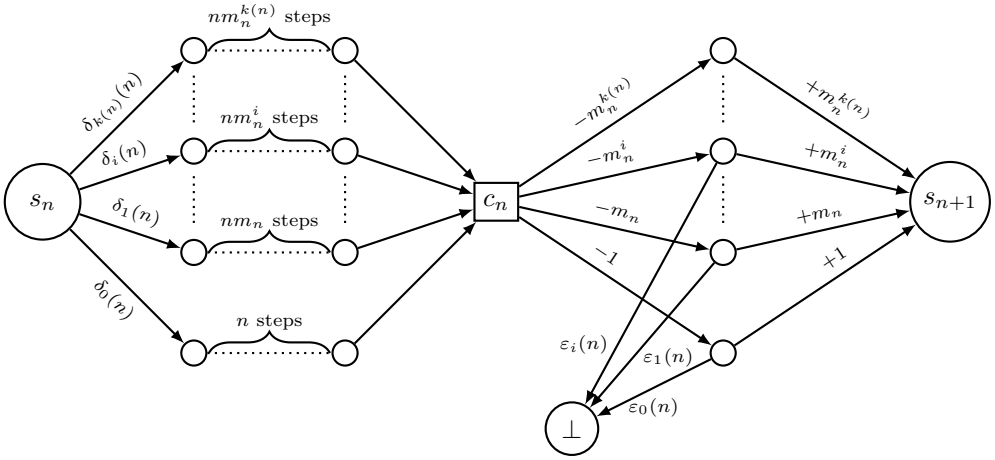

Figure 6 All transition rewards are 0 unless specified. Recall that $\sum \delta_i(n) \cdot \varepsilon_i(n)$ is convergent and $\sum \delta_j(n) \cdot \varepsilon_i(n)$ is divergent for all i, j with $j > i$. The negative reward incurred before falling into the $\bot$ state is reimbursed. We do not show it in the figure for readability. In the state before s_{n+1}, if the correct transition was chosen, the mean payoff is $-1/n$. If the incorrect transition was chosen, then either the mean payoff is $< -m_n/n$, or the risk of falling into $\bot$ is too high.

C Missing proofs from Section 4

In this part we show that a reward counter plus arbitrary finite memory does not suffice for (ε-)optimal strategies for $MP_{\liminf \geq 0}$ or for infinitely branching $TP_{\liminf \geq 0}/PP_{\liminf \geq 0}$ in countable MDPs.

First we consider $MP_{\liminf \geq 0}$ by presenting an MDP adapted from Figure 1 that has the current total reward implicit in the state and show that neither ε-optimal nor almost-sure $MP_{\liminf \geq 0}$ can be achieved by FR strategies (finite memory randomized).

We use the example from Figure 6. It is very similar to Figure 1, but differs in the following ways.

- The current total reward level is implicit in each state.
- The step counter is no longer implicit in the state.
- In the random choice, instead of changing the reward levels in each choice, it is the path length that differs.
- The definition of m_n is different, it is now $m_n \stackrel{\text{def}}{=} \sum_{i=N^*}^{n-1} m_i^{k(n)}$ with $m_{N^*} \stackrel{\text{def}}{=} 1$.

We construct a finitely branching acyclic MDP $\mathcal{M}_{\text{RI}}$ (Reward Implicit) which has the total reward implicit in the state. We do so by chaining together the gadgets from Figure 6 as is shown in Figure 2.

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

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

Proof. This follows from Lemma 41 and Lemma 42. ◀

▶ **Lemma 41.** $\texttt{val}_{\mathcal{M}_{\text{RI}}, MP_{\liminf \geq 0}}((s_0, 0)) = 1$.

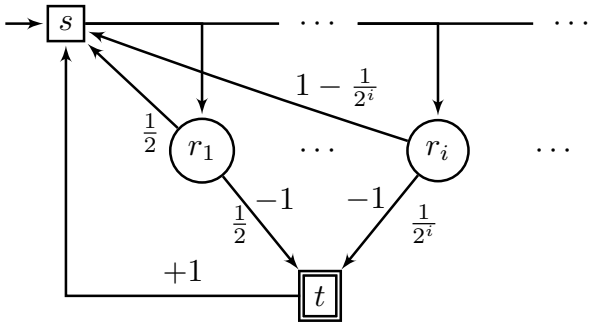

Figure 7 We present an infinitely branching MDP adapted from [16, Figure 3] and augmented with a reward structure. All of the edges carry reward 0 except the edges entering t that carry reward -1 and the edge from t to s carries reward $+1$. As a result, entering t necessarily brings the total reward down to -1 before resetting it to 0. We use a reduction to co-Büchi to show that infinite memory is required for almost-sure as well as ε-optimal strategies for $TP_{\liminf \geq 0}$ as well as $PP_{\liminf \geq 0}$.

Proof. We define a strategy σ which, in c_n always mimics the random choice in s_n. Playing according to σ, the only way to lose is by dropping into the bottom state. This is because by mimicking, the mean payoff in each gadget is lower bounded by $-1/n$. The rest of the proof is identical to Lemma 6. ◀

▶ **Lemma 42.** *Any FR strategy σ in $\mathcal{M}_{\mathrm{RI}}$ is such that $\mathcal{P}_{\mathcal{M}_{\mathrm{RI}},s_0,\sigma}(MP_{\liminf \geq 0}) = 0$.*

Proof. When playing with finitely many memory modes, there are two ways for a run in $\mathcal{M}_{\mathrm{RI}}$ to lose. Either it falls into a losing sink, or it never falls into a sink but its mean payoff is < -1. The proof that either of these occurs with probability 1 is the same as in Lemma 7. ◀

Now we construct the MDP $\mathcal{M}_{\mathrm{Restart}}$ by chaining together the gadgets from Figure 6 in the way shown in Figure 3.

▶ **Theorem 16.** *There exists a countable, finitely branching and acyclic MDP $\mathcal{M}_{\mathrm{Restart}}$ whose total reward 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 $\mathcal{P}_{\mathcal{M}_{\mathrm{Restart}},s_0,\sigma}(MP_{\liminf \geq 0}) = 0$.*

Proof. This follows from Lemma 43 and Lemma 44. ◀

▶ **Lemma 43.** *There exists a strategy σ such that $\mathcal{P}_{\mathcal{M}_{\mathrm{Restart}},s_0,\sigma}(MP_{\liminf \geq 0}) = 1$.*

Proof. The proof is identical to that of Lemma 11. ◀

▶ **Lemma 44.** *For any FR strategy σ, $\mathcal{P}_{\mathcal{M}_{\mathrm{Restart}},s_0,\sigma}(MP_{\liminf \geq 0}) = 0$.*

Proof. The proof is identical to that of Lemma 12. ◀

▶ **Theorem 17.** *There exists an infinitely branching MDP $\mathcal{M}$ as in Figure 7 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.

Proof. This follows directly from [16, Theorem 4] and the observation that in Figure 7, $TP_{\liminf \geq 0}$, $PP_{\liminf \geq 0}$ and co-Büchi objectives coincide. ◀

Consequently, when the MDP $\mathcal{M}$ is infinitely branching and has the reward counter implicit in the state, both $TP_{\liminf \geq 0}$ and $PP_{\liminf \geq 0}$ require at least a step counter.

D Missing proofs from Section 5

▶ **Definition 45.** *Let $\mathcal{M}$ be an MDP. From a given initial state s_0, the reward level in each state $s \in S$ can be any of the countably many values $r_1, r_2, \ldots$ corresponding to the rewards accumulated along all the possible paths leading to s from s_0. We then construct the MDP $R(\mathcal{M}) \stackrel{def}{=} (S', S'_\Box, S'_\bigcirc, \longrightarrow_{R(\mathcal{M})}, P')$ as follows:*

- *The state space of $R(\mathcal{M})$ is $S' \stackrel{def}{=} \{(s,r) \mid s \in S \text{ and } r \in \mathbb{R} \text{ is a reward level attainable at } s\}$. Note that S' is countable. We write s'_0 for the initial state $(s_0, 0)$.*
- *$S'_\Box \stackrel{def}{=} \{(s,r) \in S' \mid s \in S_\Box\}$ and $S'_\bigcirc \stackrel{def}{=} S' \setminus S'_\Box$.*
- *The set of transitions in $R(\mathcal{M})$ is*

$$\longrightarrow_{R(\mathcal{M})} \stackrel{def}{=} \{\, ((s,r),(s',r')) \mid (s,r),(s',r') \in S', \\ s \longrightarrow s' \text{ in } \mathcal{M} \text{ and } r' \stackrel{def}{=} r + r(s \to s')\}.$$

- *$P' : S'_\bigcirc \to \mathcal{D}(S')$ is defined such that*

$$P'(s,r)(s',r') \stackrel{def}{=} \begin{cases} P(s)(s') & \text{if } (s,r) \longrightarrow_{R(\mathcal{M})} (s',r') \\ 0 & \text{otherwise} \end{cases}$$

- *The reward for taking transition $(s,r) \longrightarrow (s',r')$ is r'.*

▶ **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.*

Proof. Let σ' be an MD (resp. Markov) strategy in $R(\mathcal{M})$ attaining $c \in [0,1]$ for $PP_{\liminf \geq 0}$ from $(s_0, 0)$. We define a strategy σ on $\mathcal{M}$ from s_0 that uses the same memory as σ' plus a reward counter. Then σ plays on $\mathcal{M}$ exactly like σ' plays on $R(\mathcal{M})$, keeping the reward counter in its memory instead of in the state. I.e., at a given state s (and step counter value m, in case σ' was a Markov strategy) and reward level r, σ plays exactly as σ' plays in state (s,r) (and step counter value m, in case σ' was a Markov strategy). By our construction of $R(\mathcal{M})$ and the definition of σ, the sequences of point rewards seen by σ' in runs on $R(\mathcal{M})$ coincide with the sequences of total rewards seen by σ in runs in $\mathcal{M}$. Hence we obtain $\mathcal{P}_{R(\mathcal{M}),(s_0,0),\sigma'}(PP_{\liminf \geq 0}) = \mathcal{P}_{\mathcal{M},s_0,\sigma}(TP_{\liminf \geq 0})$ as required. ◀

▶ **Definition 46.** *Given an MDP $\mathcal{M}$ with initial state s_0, we define the new MDP $A(\mathcal{M})$. From the initial state s_0, the reward level in each state $s \in S$ can be any of the countably many values $r_1, r_2, \ldots$ corresponding to the rewards accumulated along all the possible paths leading to s from s_0.*

We then construct $A(\mathcal{M}) \stackrel{def}{=} (S', S'_\Box, S'_\bigcirc, \longrightarrow_{A(\mathcal{M})}, P')$ as follows:

- *The state space of $A(\mathcal{M})$ is*

$$S' \stackrel{def}{=} \{(s,n,r) \mid s \in S, n \in \mathbb{N} \text{ and } r \in \mathbb{R} \text{ is a reward level attainable at } s \text{ at step } n\}$$

Note that S' is countable. We write s'_0 for the initial state $(s_0, 0, 0)$ of $A(\mathcal{M})$.
- *$S'_\Box \stackrel{def}{=} \{(s,n,r) \in S' \mid s \in S_\Box\}$ and $S'_\bigcirc \stackrel{def}{=} S' \setminus S'_\Box$.*
- *The set of transitions in $A(\mathcal{M})$ is*

$$\longrightarrow_{A(\mathcal{M})} \stackrel{def}{=} \{\, ((s,n,r),(s',n+1,r')) \mid \\ (s,n,r),(s',n+1,r') \in S', \\ s \longrightarrow s' \text{ in } \mathcal{M} \text{ and } r' = r + r(s \to s')\}.$$

- $P' : S'_{\bigcirc} \to \mathcal{D}(S')$ *is defined such that*

$$P'(s,n,r)(s',n',r') \stackrel{def}{=} \begin{cases} P(s)(s') & \text{if } (s,n,r) \to_{A(\mathcal{M})} (s',n',r') \\ 0 & \text{otherwise} \end{cases}$$

- *The reward for taking transition* $(s,n,r) \longrightarrow (s',n',r')$ *is* r'/n'.

D.1 Proofs from Section 5.1

In this section we consider finitely branching MDPs. We need the following technical lemma that holds only for finitely branching MDPs.

▶ **Lemma 47.** *Given a finitely branching countable MDP* $\mathcal{M}$*, a subset* $T \subseteq \to$ *of the transitions and a state* s*, we have*

$$\mathtt{val}_{\mathcal{M},\neg \mathsf{F} T}(s) < 1 \;\Rightarrow\; \exists k \in \mathbb{N}.\, \mathtt{val}_{\mathcal{M},\neg \mathsf{F}^{\leq k} T}(s) < 1$$

i.e., if it is impossible to completely avoid T *then there is a bounded threshold* k *and a fixed nonzero chance of seeing* T *within* $\leq k$ *steps, regardless of the strategy.*

Proof. If suffices to show that $\forall k \in \mathbb{N}.\, \mathtt{val}_{\mathcal{M},\neg \mathsf{F}^{\leq k} T}(s) = 1$ implies $\mathtt{val}_{\mathcal{M},\neg \mathsf{F} T}(s) = 1$. Since $\mathcal{M}$ is finitely branching, the state s has only finitely many successors $\{s_1, \ldots, s_n\}$.

Consider the case where s is a controlled state. If we had the property $\forall 1 \leq i \leq n\, \exists k_i \in \mathbb{N}.\, \mathtt{val}_{\mathcal{M},\neg \mathsf{F}^{\leq k_i} T}(s_i) < 1$ then we would have $\mathtt{val}_{\mathcal{M},\neg \mathsf{F}^{\leq k} T}(s) < 1$ for $k = (\max_{1 \leq i \leq n} k_i) + 1$ which contradicts our assumption. Thus there must exist an $i \in \{1, \ldots, n\}$ with $\forall k \in \mathbb{N}.\, \mathtt{val}_{\mathcal{M},\neg \mathsf{F}^{\leq k} T}(s_i) = 1$. We define a strategy σ that chooses the successor state s_i when in state s.

Similarly, if s is a random state, we must have $\forall k \in \mathbb{N}.\, \mathtt{val}_{\mathcal{M},\neg \mathsf{F}^{\leq k} T}(s_i) = 1$ for all its successors s_i.

By using our constructed strategy σ, we obtain $\mathcal{P}_{\mathcal{M},s,\sigma}(\neg \mathsf{F} T) = 1$ and thus $\mathtt{val}_{\mathcal{M},\neg \mathsf{F} T}(s) = 1$ as required. ◀

▷ Claim 28.

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

Proof.

$$
\begin{aligned}
&\mathcal{P}_{\mathcal{M}',s_0,\sigma}\left(\bigcap_{i\in\mathbb{N}} \mathrm{Safety}_i^{n_i}\right)\\
&\geq \mathcal{P}_{\mathcal{M}',s_0,\sigma}\left(\bigcap_{k\in\mathbb{N}} \mathsf{F}(\mathrm{Safety}_k) \cap \bigcap_{i\in\mathbb{N}} \mathrm{Safety}_i^{n_i}\right)\\
&= \mathcal{P}_{\mathcal{M}',s_0,\sigma}\left(\left(\bigcap_{k\in\mathbb{N}} \mathsf{F}(\mathrm{Safety}_k) \cap \bigcap_{i\in\mathbb{N}} \mathrm{Safety}_i^{n_i}\right) \cup \left(\overline{\bigcap_{k\in\mathbb{N}} \mathsf{F}(\mathrm{Safety}_k)} \cap \bigcap_{k\in\mathbb{N}} \mathsf{F}(\mathrm{Safety}_k)\right)\right)\\
&= \mathcal{P}_{\mathcal{M}',s_0,\sigma}\left(\bigcap_{k\in\mathbb{N}} \mathsf{F}(\mathrm{Safety}_k) \cap \left(\bigcap_{i\in\mathbb{N}} \mathrm{Safety}_i^{n_i} \cup \overline{\bigcap_{k\in\mathbb{N}} \mathsf{F}(\mathrm{Safety}_k)}\right)\right)\\
&= 1 - \mathcal{P}_{\mathcal{M}',s_0,\sigma}\left(\overline{\bigcap_{k\in\mathbb{N}} \mathsf{F}(\mathrm{Safety}_k)} \cup \left(\overline{\bigcap_{i\in\mathbb{N}} \mathrm{Safety}_i^{n_i}} \cap \bigcap_{k\in\mathbb{N}} \mathsf{F}(\mathrm{Safety}_k)\right)\right)\\
&\geq 1 - \mathcal{P}_{\mathcal{M}',s_0,\sigma}\left(\overline{\bigcap_{k\in\mathbb{N}} \mathsf{F}(\mathrm{Safety}_k)}\right) - \mathcal{P}_{\mathcal{M}',s_0,\sigma}\left(\overline{\bigcap_{i\in\mathbb{N}} \mathrm{Safety}_i^{n_i}} \cap \bigcap_{k\in\mathbb{N}} \mathsf{F}(\mathrm{Safety}_k)\right)\\
&= \mathcal{P}_{\mathcal{M}',s_0,\sigma}\left(PP_{\liminf\geq 0}\right) - \mathcal{P}_{\mathcal{M}',s_0,\sigma}\left(\overline{\bigcap_{i\in\mathbb{N}} \mathrm{Safety}_i^{n_i}} \cap \bigcap_{k\in\mathbb{N}} \mathsf{F}(\mathrm{Safety}_k)\right) && \text{by (3)}\\
&\geq \mathtt{val}_{\mathcal{M}',PP_{\liminf\geq 0}}(s_0) - \varepsilon - \mathcal{P}_{\mathcal{M}',s_0,\sigma}\left(\bigcup_{i\in\mathbb{N}} \overline{\mathrm{Safety}_i^{n_i}} \cap \bigcap_{k\in\mathbb{N}} \mathsf{F}(\mathrm{Safety}_k)\right) && \text{by (2)}\\
&\geq \mathtt{val}_{\mathcal{M}',PP_{\liminf\geq 0}}(s_0) - \varepsilon - \sum_{i\in\mathbb{N}} \mathcal{P}_{\mathcal{M}',s_0,\sigma}\left(\overline{\mathrm{Safety}_i^{n_i}} \cap \bigcap_{k\in\mathbb{N}} \mathsf{F}(\mathrm{Safety}_k)\right)\\
&\geq \mathtt{val}_{\mathcal{M}',PP_{\liminf\geq 0}}(s_0) - \varepsilon - \sum_{i\in\mathbb{N}} \varepsilon_i && \text{by (5)}\\
&= \mathtt{val}_{\mathcal{M}',PP_{\liminf\geq 0}}(s_0) - 2\varepsilon
\end{aligned}
$$

◀

▷ Claim 29.

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

Proof. First we show that

$$\mathtt{Transience} \subseteq PP_{\liminf\geq 0} \quad \text{in } \mathcal{M}''. \tag{12}$$

Let $\rho \in \mathtt{Transience}$ be a transient run. Then ρ can never visit the state $\bot$. Moreover, ρ must eventually leave every finite set forever. In particular ρ must satisfy $\mathsf{FG}(\neg \mathrm{Bubble}_{n_i}(s_0))$ for every i, since $\mathrm{Bubble}_{n_i}(s_0)$ is finite, because $\mathcal{M}''$ is finitely branching. Thus ρ must either fall into G_{safe}, in which case it satisfies $PP_{\liminf\geq 0}$, or for every i, ρ must eventually leave $\mathrm{Bubble}_{n_i}(s_0)$ forever. By definition of $\mathrm{Bubble}_{n_i}(s_0)$ and $\mathcal{M}''$, the run ρ must eventually stop seeing rewards $< -2^{-i}$ for every i. In this case ρ also satisfies $PP_{\liminf\geq 0}$. Thus (12).

Secondly, we show that

$$\forall \sigma''.\ \mathcal{P}_{\mathcal{M}'',s_0,\sigma''}(PP_{\liminf\geq 0} \cap \overline{\mathtt{Transience}}) = 0. \tag{13}$$

i.e., except for a null-set, $PP_{\liminf\geq 0}$ implies $\mathtt{Transience}$ in $\mathcal{M}''$.

Let σ'' be an arbitrary strategy from s_0 in $\mathcal{M}''$ and $\mathfrak{R}$ be the set of all runs induced by it. For every $s \in S$, let $\mathfrak{R}_s \stackrel{\text{def}}{=} \{\rho \in \mathfrak{R} \mid \rho \text{ satisfies } \mathsf{GF}(s)\}$ be the set of runs seeing state s infinitely often. In particular, any run $\rho \in \mathfrak{R}_s$ is not transient. Indeed, $\overline{\texttt{Transience}} = \bigcup_{s \in S} \mathfrak{R}_s$. We want to show that for every state $s \in S$ and strategy σ''

$$\mathcal{P}_{\mathcal{M}'',s_0,\sigma''}(PP_{\liminf \geq 0} \cap \mathfrak{R}_s) = 0. \tag{14}$$

Since any runs seeing a state in G_{safe} are transient, any $\mathfrak{R}_s$ with $s \in G_{\text{safe}}$ must be empty. Similarly, any run seeing $\bot$ is losing for $PP_{\liminf \geq 0}$ by construction. Hence we have (14) for any state s where $s = \bot$ or $s \in G_{\text{safe}}$.

Now consider $\mathfrak{R}_s$ where s is neither in G_{safe} nor $\bot$. Let $T_{neg} \stackrel{\text{def}}{=} \{t \in \longrightarrow \mid r(t) < 0\}$ be the subset of transitions with negative rewards in $\mathcal{M}''$.

We now show that $\texttt{val}_{\mathcal{M}'',\neg \mathsf{F} T_{neg}}(s) < 1$ by assuming the opposite and deriving a contradiction. Assume that $\texttt{val}_{\mathcal{M}'',\neg \mathsf{F} T_{neg}}(s) = 1$. The objective $\neg \mathsf{F} T_{neg}$ is a safety objective. Thus, since $\mathcal{M}''$ is finitely branching, there exists a strategy from s that surely avoids T_{neg} (always pick an optimal move) [18, 16]. (This does not hold in infinitely branching MDPs where optimal moves might not exist.) However, by construction of $\mathcal{M}''$, this implies that $s \in G_{\text{safe}}$. Contradiction. Thus $\texttt{val}_{\mathcal{M}'',\neg \mathsf{F} T_{neg}}(s) < 1$.

Since $\mathcal{M}''$ is finitely branching, we can apply Lemma 47 and obtain that there exists a threshold k_s such that $\texttt{val}_{\mathcal{M}'',\neg \mathsf{F}^{\leq k_s} T_{neg}}(s) < 1$. Therefore $\delta_s \stackrel{\text{def}}{=} 1 - \texttt{val}_{\mathcal{M}'',\neg \mathsf{F}^{\leq k_s} T_{neg}}(s) > 0$. Thus, under every strategy, upon visiting s there is a chance $\geq \delta_s$ of seeing a transition in T_{neg} within the next $\leq k_s$ steps. Moreover, the subset $T^s_{neg} \subseteq T_{neg}$ of transitions that can be reached in $\leq k_s$ steps from s is finite, since $\mathcal{M}''$ is finitely branching. So the maximum of the rewards in T^s_{neg} is still negative, i.e., $\ell_s \stackrel{\text{def}}{=} \max\{r(t) \mid t \in T^s_{neg}\} < 0$. Let $T_{\leq \ell} \stackrel{\text{def}}{=} \{t \in \longrightarrow \mid r(t) \leq \ell_s\}$ be the subset of transitions with rewards $\leq \ell_s$ in $\mathcal{M}''$.

Thus, under *every* strategy, upon visiting s there is a chance $\geq \delta_s$ of seeing a transition in $T_{\leq \ell}$ within the next $\leq k_s$ steps.

Define $\mathfrak{R}^i_s \stackrel{\text{def}}{=} \{\rho \in \mathfrak{R} \mid \rho \text{ sees } s \text{ at least } i \text{ times}\}$, so we get $\mathfrak{R}_s = \bigcap_{i \in \mathbb{N}} \mathfrak{R}^i_s$. We obtain

$$\begin{aligned}
&\sup_{\sigma''} \mathcal{P}_{\mathcal{M}'',s_0,\sigma''}(PP_{\liminf \geq 0} \cap \mathfrak{R}_s) \\
&\leq \sup_{\sigma''} \mathcal{P}_{\mathcal{M}'',s_0,\sigma''}(\mathsf{FG} \neg T_{\leq \ell} \cap \mathfrak{R}_s) && \text{set inclusion} \\
&= \sup_{\sigma''} \lim_{n \to \infty} \mathcal{P}_{\mathcal{M}'',s_0,\sigma''}(\mathsf{F}^{\leq n} \mathsf{G} \neg T_{\leq \ell} \cap \mathfrak{R}_s) && \text{continuity of measures} \\
&\leq \sup_{\sigma'''} \mathcal{P}_{\mathcal{M}'',s,\sigma'''}(\mathsf{G} \neg T_{\leq \ell} \cap \mathfrak{R}_s) && s \text{ visited after } > n \text{ steps} \\
&= \sup_{\sigma'''} \mathcal{P}_{\mathcal{M}'',s,\sigma'''}(\mathsf{G} \neg T_{\leq \ell} \cap \bigcap_{i \in \mathbb{N}} \mathfrak{R}^i_s) && \text{def. of } \mathfrak{R}^i_s \\
&= \sup_{\sigma'''} \lim_{i \to \infty} \mathcal{P}_{\mathcal{M}'',s,\sigma'''}(\mathsf{G} \neg T_{\leq \ell} \cap \mathfrak{R}^i_s) && \text{continuity of measures} \\
&\leq \lim_{i \to \infty} (1 - \delta_s)^i = 0 && \text{by def. of } \mathfrak{R}^i_s \text{ and } \delta_s
\end{aligned}$$

and thus (14).

From this we obtain $\mathcal{P}_{\mathcal{M}'',s_0,\sigma''}(PP_{\liminf \geq 0} \cap \overline{\texttt{Transience}}) = \mathcal{P}_{\mathcal{M}'',s_0,\sigma''}(PP_{\liminf \geq 0} \cap \bigcup_{s \in S} \mathfrak{R}_s) \leq \sum_{s \in S} \mathcal{P}_{\mathcal{M}'',s_0,\sigma''}(PP_{\liminf \geq 0} \cap \mathfrak{R}_s) = 0$ and thus (13).

From (12) and (13) we obtain that for every σ'' we have

$$\begin{aligned}
&\mathcal{P}_{\mathcal{M}'',s_0,\sigma''}(PP_{\liminf\geq 0}) \\
&= \mathcal{P}_{\mathcal{M}'',s_0,\sigma''}(PP_{\liminf\geq 0} \cap \texttt{Transience}) + \mathcal{P}_{\mathcal{M}'',s_0,\sigma''}(PP_{\liminf\geq 0} \cap \overline{\texttt{Transience}}) \\
&= \mathcal{P}_{\mathcal{M}'',s_0,\sigma''}(\texttt{Transience}) + 0 \\
&= \mathcal{P}_{\mathcal{M}'',s_0,\sigma''}(\texttt{Transience})
\end{aligned}$$

and thus Claim 29. ◀

D.2 Proofs from Section 5.2

In this section we consider infinitely branching MDPs. In the following theorem we show how to obtain ε-optimal deterministic Markov strategies for $PP_{\liminf\geq 0}$. We do this by deriving ε-optimal MD strategies in $S(\mathcal{M})$ via a reduction to a safety objective.

▶ **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 $\mathcal{M}$.*

Proof. Let $\varepsilon > 0$. We work in $S(\mathcal{M})$ by encoding the step counter into the states of $\mathcal{M}$. Thus $S(\mathcal{M})$ is an acyclic MDP with implicit step counter and corresponding initial state $s_0' = (s_0, 0)$.

We consider a general (not necessarily MD) ε-optimal strategy σ for $PP_{\liminf\geq 0}$ from s_0' on $S(\mathcal{M})$, i.e.,

$$\mathcal{P}_{S(\mathcal{M}),s_0',\sigma}(PP_{\liminf\geq 0}) \geq \texttt{val}_{S(\mathcal{M}),PP_{\liminf\geq 0}}(s_0') - \varepsilon. \tag{15}$$

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

$$PP_{\liminf\geq 0} = \bigcap_{i\in\mathbb{N}} \mathsf{F}(\text{Safety}_i) \tag{16}$$

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

$$\mathsf{F}(\text{Safety}_i) = \bigcup_{k\in\mathbb{N}} \text{Safety}_i^k. \tag{17}$$

By continuity of measures from above we get

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

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

$$\mathcal{P}_{S(\mathcal{M}),s_0',\sigma}\left(\mathsf{F}(\text{Safety}_i) \cap \overline{\text{Safety}_i^{n_i}}\right) \leq \varepsilon_i. \tag{18}$$

Now we can show the following claim.

▷ Claim 48.

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

Proof.

$$\begin{aligned}
&\mathcal{P}_{S(\mathcal{M}),s_0',\sigma}\left(\bigcap_{i\in\mathbb{N}}\text{Safety}_i^{n_i}\right)\\
&\geq \mathcal{P}_{S(\mathcal{M}),s_0',\sigma}\left(\bigcap_{k\in\mathbb{N}}\mathsf{F}(\text{Safety}_k)\cap\bigcap_{i\in\mathbb{N}}\text{Safety}_i^{n_i}\right)\\
&= \mathcal{P}_{S(\mathcal{M}),s_0',\sigma}\left(\left(\bigcap_{k\in\mathbb{N}}\mathsf{F}(\text{Safety}_k)\cap\bigcap_{i\in\mathbb{N}}\text{Safety}_i^{n_i}\right)\cup\left(\overline{\bigcap_{k\in\mathbb{N}}\mathsf{F}(\text{Safety}_k)}\cap\bigcap_{k\in\mathbb{N}}\mathsf{F}(\text{Safety}_k)\right)\right)\\
&= \mathcal{P}_{S(\mathcal{M}),s_0',\sigma}\left(\bigcap_{k\in\mathbb{N}}\mathsf{F}(\text{Safety}_k)\cap\left(\bigcap_{i\in\mathbb{N}}\text{Safety}_i^{n_i}\cup\overline{\bigcap_{k\in\mathbb{N}}\mathsf{F}(\text{Safety}_k)}\right)\right)\\
&= 1-\mathcal{P}_{S(\mathcal{M}),s_0',\sigma}\left(\overline{\bigcap_{k\in\mathbb{N}}\mathsf{F}(\text{Safety}_k)}\cup\left(\overline{\bigcap_{i\in\mathbb{N}}\text{Safety}_i^{n_i}}\cap\bigcap_{k\in\mathbb{N}}\mathsf{F}(\text{Safety}_k)\right)\right)\\
&\geq 1-\mathcal{P}_{S(\mathcal{M}),s_0',\sigma}\left(\overline{\bigcap_{k\in\mathbb{N}}\mathsf{F}(\text{Safety}_k)}\right)-\mathcal{P}_{S(\mathcal{M}),s_0',\sigma}\left(\overline{\bigcap_{i\in\mathbb{N}}\text{Safety}_i^{n_i}}\cap\bigcap_{k\in\mathbb{N}}\mathsf{F}(\text{Safety}_k)\right)\\
&= \mathcal{P}_{S(\mathcal{M}),s_0',\sigma}(PP_{\liminf\geq 0})-\mathcal{P}_{S(\mathcal{M}),s_0',\sigma}\left(\overline{\bigcap_{i\in\mathbb{N}}\text{Safety}_i^{n_i}}\cap\bigcap_{k\in\mathbb{N}}\mathsf{F}(\text{Safety}_k)\right) && \text{by (16)}\\
&\geq \mathtt{val}_{S(\mathcal{M}),PP_{\liminf\geq 0}}(s_0')-\varepsilon-\mathcal{P}_{S(\mathcal{M}),s_0',\sigma}\left(\bigcup_{i\in\mathbb{N}}\overline{\text{Safety}_i^{n_i}}\cap\bigcap_{k\in\mathbb{N}}\mathsf{F}(\text{Safety}_k)\right) && \text{by (15)}\\
&\geq \mathtt{val}_{S(\mathcal{M}),PP_{\liminf\geq 0}}(s_0')-\varepsilon-\sum_{i\in\mathbb{N}}\mathcal{P}_{S(\mathcal{M}),s_0',\sigma}\left(\overline{\text{Safety}_i^{n_i}}\cap\bigcap_{k\in\mathbb{N}}\mathsf{F}(\text{Safety}_k)\right)\\
&\geq \mathtt{val}_{S(\mathcal{M}),PP_{\liminf\geq 0}}(s_0')-\varepsilon-\sum_{i\in\mathbb{N}}\varepsilon_i && \text{by (18)}\\
&= \mathtt{val}_{S(\mathcal{M}),PP_{\liminf\geq 0}}(s_0')-2\varepsilon
\end{aligned}$$

◀

Let $\varphi \stackrel{\text{def}}{=} \bigcap_{i\in\mathbb{N}}\text{Safety}_i^{n_i} \subseteq PP_{\liminf\geq 0}$. It follows from Claim 48 that

$$\mathtt{val}_{S(\mathcal{M}),\varphi}(s_0') \geq \mathtt{val}_{S(\mathcal{M}),PP_{\liminf\geq 0}}(s_0') - 2\varepsilon. \tag{19}$$

The objective φ is a safety objective on $S(\mathcal{M})$. Therefore, since $S(\mathcal{M})$ is acyclic, we can apply Lemma 24 to obtain a uniformly ε-optimal MD strategy σ' for φ. Thus

$$\begin{aligned}
&\mathcal{P}_{S(\mathcal{M}),s_0',\sigma'}(PP_{\liminf\geq 0})\\
&\geq \mathcal{P}_{S(\mathcal{M}),s_0',\sigma'}(\varphi) && \text{set inclusion}\\
&\geq \mathtt{val}_{S(\mathcal{M}),\varphi}(s_0')-\varepsilon && \sigma' \text{ is } \varepsilon\text{-opt.}\\
&\geq \mathtt{val}_{S(\mathcal{M}),PP_{\liminf\geq 0}}(s_0')-3\varepsilon. && \text{by (19)}
\end{aligned}$$

Thus σ' is a 3ε-optimal MD strategy for $PP_{\liminf\geq 0}$ in $S(\mathcal{M})$.

By Remark 21 this then yields a 3ε-optimal Markov strategy for $PP_{\liminf\geq 0}$ from s_0 in $\mathcal{M}$, since runs in $\mathcal{M}$ and $S(\mathcal{M})$ coincide wrt. $PP_{\liminf\geq 0}$. ◀

▶ **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.*

Proof. We consider the encoded system $A(\mathcal{M})$ in which both step counter and reward counter are implicit in the state. Recall that the partial mean payoffs in $\mathcal{M}$ correspond exactly to point rewards in $A(\mathcal{M})$. Since $A(\mathcal{M})$ has an encoded step counter, Theorem 32 gives us ε-optimal MD strategies for $PP_{\liminf \geq 0}$ in $A(\mathcal{M})$. Lemma 23 allows us to translate these strategies back to $\mathcal{M}$ with a memory overhead of just a reward counter and a step counter as required. ◀

▶ **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.*

Proof. We consider the encoded system $R(\mathcal{M})$ in which the reward counter is implicit in the state. Recall that total rewards in $\mathcal{M}$ correspond exactly to point rewards in $R(\mathcal{M})$. We then apply Theorem 32 to $R(\mathcal{M})$ to obtain ε-optimal MD strategies for $PP_{\liminf \geq 0}$ in $S(R(\mathcal{M}))$. Remark 21 allows us to translate these MD strategies back to $R(\mathcal{M})$ with a memory overhead of just a step counter. Then we apply Lemma 20 to translate these Markov strategies back to $\mathcal{M}$ with a memory overhead of just a reward counter. Hence ε-optimal strategies for $TP_{\liminf \geq 0}$ in $\mathcal{M}$ just use a step counter and a reward counter as required. ◀

▶ Remark 49. While ε-optimal strategies for mean payoff and total payoff (in infinitely branching MDPs) have the same memory requirements, the step counter and the reward counter do not arise in the same way. Both the step counter and reward counter used in ε-optimal strategies for mean payoff arise from the construction of $A(\mathcal{M})$. However, in the case for total payoff, only the reward counter arises from the construction of $R(\mathcal{M})$. The step counter on the other hand arises from the Markov strategy needed for point payoff in $R(\mathcal{M})$.

▶ **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.*

Proof. To obtain the result for $PP_{\liminf \geq 0}$, we work in $S(\mathcal{M})$ and we apply Theorem 32 to obtain ε-optimal MD strategies from every state of $S(\mathcal{M})$. Since $PP_{\liminf \geq 0}$ is a tail objective, Theorem 25 yields an MD strategy that is optimal from every state of $S(\mathcal{M})$ that has an optimal strategy. By Remark 21 we can translate this MD strategy on $S(\mathcal{M})$ back to a Markov strategy in $\mathcal{M}$, which is optimal for $PP_{\liminf \geq 0}$ from s_0 (provided that s_0 admits any optimal strategy at all).

Consider the case for $MP_{\liminf \geq 0}$. First we place ourselves in $A(\mathcal{M})$ and apply Theorem 32 to obtain ε-optimal MD strategies from every state of $A(\mathcal{M})$. From Theorem 25 we obtain a single MD strategy that is optimal from every state of $A(\mathcal{M})$ that has an optimal strategy. By Lemma 23 we can translate this MD strategy on $A(\mathcal{M})$ back to a strategy on $\mathcal{M}$ with a step counter and a reward counter. Provided that s_0 admits any optimal strategy at all, we obtain an optimal strategy for $MP_{\liminf \geq 0}$ from s_0 that uses only a step counter and a reward counter.

The case for $TP_{\liminf \geq 0}$ is similar. We place ourselves in $S(R(\mathcal{M}))$ and apply Corollary 34 to obtain ε-optimal MD strategies for $TP_{\liminf \geq 0}$ from every state of $S(R(\mathcal{M}))$. While $TP_{\liminf \geq 0}$ is not tail in $\mathcal{M}$, it is tail in $S(R(\mathcal{M}))$, and thus we can apply Theorem 25 to

obtain a single MD strategy that is optimal from every state of $S(R(\mathcal{M}))$ that has an optimal strategy. The result then follows from Lemma 20 and Remark 21. ◀

E Strengthening results

We show that the counterexamples presented in Section 3 can be modified s.t. all transition rewards are either -1, 0, or $+1$ and the maximal degree of branching is 2. I.e., the hardness does not depend on arbitrarily large rewards or degrees of branching.

Consider a new MDP $\mathcal{M}$ based on the MDP constructed in Figure 1 which now undergoes the following changes. The rewards on transitions are now limited to -1, 0 or 1. To compensate for the smaller rewards, in the n-th gadget, each transition bearing a reward is replaced by $k(n) \cdot m_n$ transitions as follows. If the original transition had reward $j \cdot m_n$ then that transition is replaced with $j \cdot m_n$ transitions with reward 1, and $(k(n) - j) \cdot m_n$ transitions with reward 0. Symmetrically all negatively weighted transitions are similarly replaced by transitions with rewards -1 and 0.

We further alter $\mathcal{M}$ by modifying Figure 1 such that the branching degree is bounded by 2. We do this by replacing the outgoing transitions in states s_n and c_n of each gadget by binary trees with accordingly adjusted probabilities such that there is still a probability of $\delta_i(n)$ of receiving reward $i \cdot m_n$ in each gadget for $i \in \{0, 1, ..., k(n)\}$.

To adjust for the increased path lengths incurred by the modifications to each gadget, the construction in Figure 2 is accordingly modified by padding each vertical column of white states with extra transitions based on the number of transitions present in the matching gadget. As a result, path length is preserved even when skipping gadgets. The construction in Figure 3 is similarly modified.

This construction allows us to obtain the following properties.

▶ Remark 50. There exists a countable, acyclic MDP $\mathcal{M}$, whose step counter is implicit in the state, whose rewards on transitions are in $\{-1, 0, 1\}$ and whose branching degree is bounded by 2 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 step counter plus finite memory strategies for any $\varepsilon < 1$ for the $MP_{\liminf \geq 0}$ objective for countable MDPs.

This follows from Lemma 6, Lemma 7 and the above construction.

▶ Remark 51. There exists a countable, acyclic MDP $\mathcal{M}$, whose step counter is implicit in the state, whose rewards on transitions are in $\{-1, 0, 1\}$ and whose branching degree is bounded by 2 for which s_0 is almost surely winning and any 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 with a step counter plus finite memory for countable MDPs.

This follows from Lemma 11, Lemma 12 and the above construction.

▶ Remark 52. The two previous remarks also hold for the $TP_{\liminf \geq 0}$ objective with no modifications to their respective constructions or their proofs.

▶ Remark 53. The result from Lemma 7 holds even for strategies σ whose memory grows unboundedly, but slower than $k(n) - 1$. That is to say that there exists a countable, acyclic MDP $\mathcal{M}$, whose step counter is implicit in the state such that $\mathtt{val}_{\mathcal{M}, MP_{\liminf \geq 0}}(s_0) = 1$ and any strategy σ with memory $< k(n) - 1$ is such that $\mathcal{P}_{\mathcal{M}, \sigma, s_0}(MP_{\liminf \geq 0}) = 0$. The result then follows since in every gadget at least one memory mode will confuse at least two states $i(n), j(n) : \mathbb{N} \to \{0, 1, ..., k(n) - 1\}$.

The results from Section 4 can similarly be strengthened. Consider the construction in Figure 6. In the random choice, the transition rewards are already all $+1$, so only the branching degree needs to be adjusted by padding the choice with a binary tree as above. In the controlled choice, the transitions carrying reward $\pm m_n^i$ are replaced by m_n^i transitions

each bearing reward ± 1 respectively. Therefore, the path lengths increase in the following way in the n-th gadget. In s_n and c_n, the binary trees increase path length by up to $\lceil \lg(k(n)+1) \rceil$ (where lg is the logarithm to base 2) and after c_n the path length increases by up to $m_n^{k(n)}$ twice.

Consider the scenario where the play took the i-th random choice and the player makes the 'best' mistake where they choose transition $i+1$. We show that, even in this best error case (and thus in all other error cases), the newly added path lengths do still not help to prevent seeing a mean payoff $\leq -1/2$ in the n-th gadget. In this case, in the state between c_n and s_{n+1}, the total payoff is $-m_n^{i+1}$ and the total number of steps taken by the play so far is upper bounded by

$$\beta_n \stackrel{\text{def}}{=} \left(\sum_{i=N^*}^{n-1} 2\lceil \lg(k(i)+1) \rceil + 2m_i^{k(i)} \right) + 2\lceil \lg(k(n)+1) \rceil + m_n^i + m_n^{i+1}.$$

Recall that $m_n \stackrel{\text{def}}{=} \sum_{i=N^*}^{n-1} m_i^{k(n)}$ with $m_{N^*} \stackrel{\text{def}}{=} 1$, and this is the definition of m_n from Appendix C which is different from the definition of m_n in Section 3. Note that $k(n)$ is very slowly growing, so it follows that

$$\beta_n \leq 3m_n + m_n^i + m_n^{i+1} \leq 2m_n^{i+1}.$$

That is to say that the mean payoff is $\leq \dfrac{-m_n^{i+1}}{2m_n^{i+1}} = -1/2$. As a result, in the case of a bad aggressive decision, the mean payoff will still drop below $-1/2$ in this modified MDP (instead of dropping below -1 in the original MDP). This is just as good to falsify $MP_{\liminf \geq 0}$.

Thus we obtain the following two results.

▶ Remark 54. There exists a countable, acyclic MDP $\mathcal{M}$, whose reward counter is implicit in the state, whose rewards on transitions are in $\{-1, 0, 1\}$ and whose branching degree is bounded by 2 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 step counter plus finite memory strategies for any $\varepsilon < 1$ for the $MP_{\liminf \geq 0}$ objective for countable MDPs.

This follows from Lemma 41, Lemma 42 and the above construction.

▶ Remark 55. There exists a countable, acyclic MDP $\mathcal{M}$, whose reward counter is implicit in the state, whose rewards on transitions are in $\{-1, 0, 1\}$ and whose branching degree is bounded by 2 for which s_0 is almost surely winning and any 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 with a step counter plus finite memory for countable MDPs.

This follows from Lemma 43, Lemma 44 and the above construction.