Finite-memory Strategies for Almost-sure Energy-MeanPayoff Objectives in MDPs

Mohan Dantam
School of Informatics, University of Edinburgh, UK

Richard Mayr
School of Informatics, University of Edinburgh, UK

Abstract

We consider finite-state Markov decision processes with the combined Energy-MeanPayoff objective. The controller tries to avoid running out of energy while simultaneously attaining a strictly positive mean payoff in a second dimension.

We show that *finite memory* suffices for almost surely winning strategies for the Energy-MeanPayoff objective. This is in contrast to the closely related Energy-Parity objective, where almost surely winning strategies require infinite memory in general.

We show that exponential memory is sufficient (even for deterministic strategies) and necessary (even for randomized strategies) for almost surely winning Energy-MeanPayoff. The upper bound holds even if the strictly positive mean payoff part of the objective is generalized to multidimensional strictly positive mean payoff.

Finally, it is decidable in pseudo-polynomial time whether an almost surely winning strategy exists.

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

Keywords and phrases Markov decision processes, energy, mean payoff, parity, strategy complexity

Digital Object Identifier 10.4230/LIPIcs...

1 Introduction

Background. Markov decision processes (MDPs) are a standard model for dynamic systems that exhibit both stochastic and controlled behavior [29]. MDPs play a prominent role in many domains, e.g., artificial intelligence and machine learning [33, 31], control theory [5, 1], operations research and finance [32, 24, 10, 30], and formal verification [2, 32, 20, 15, 3, 12].

An MDP is a directed graph where states are either controlled or random. If the current state is controlled then the controller can choose a distribution over all possible successor states. If the current state is random then the next state is chosen according to a fixed probability distribution. One assigns numeric rewards to transitions (and this can be generalized to multidimensional rewards). Moreover, priorities (aka colors), encoded by bounded non-negative numbers, are assigned to 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 *strategy complexity* of a given objective is the amount of memory (and randomization) needed for an optimal (resp. ε-optimal) strategy. Common cases include memoryless strategies, finite-memory strategies, Markov strategies (using a discrete clock, aka step counter), and general infinite-memory strategies.

Related work. The Parity, MeanPayoff and Energy objectives have been extensively studied in the formal verification community. A run satisfies the (min-even) *Parity objective* iff the minimal priority that appears infinitely often in the run is even. It subsumes all ω-regular objectives, and in particular safety, liveness, fairness, etc. The *MeanPayoff objective* requires that the limit average reward per transition along a run is positive (resp. non-negative in some settings). MeanPayoff objectives go back to a 1957 paper by Gillette [21] and have been widely studied, due to their relevance for efficient control. The *Energy objective* [11] requires that the accumulated reward at any time in a run stays above some finite threshold (typically 0). The intuition is that a controlled system has some finite initial energy level that must never become depleted.

Combinations of these objectives have also been studied, where the runs need to satisfy several of the above conditions simultaneously.

The existence of almost surely winning strategies for *MeanPayoff-Parity* in MDPs is decidable in polynomial time [13]. These strategies require only finite memory for MeanPayoff > 0 [22], but infinite memory for MeanPayoff ≥ 0 [14].

The existence of almost surely winning strategies for *Energy-Parity* in MDPs is decidable in $\mathsf{NP} \cap \mathsf{coNP}$ and in pseudo-polynomial time [26]. (The $\mathsf{NP} \cap \mathsf{coNP}$ upper bound holds even for turn-based stochastic games [27].) Almost surely winning strategies in MDPs require only finite memory in the special case of Energy-Büchi [13], but infinite memory for Energy-co-Büchi and thus for Energy-Parity [26]. However, ε-optimal strategies for Energy-Parity require only finite (at most doubly exponential) memory, and the value can be effectively approximated in doubly exponential time (even for turn-based stochastic games) [18].

The *Energy-MeanPayoff* objective is similar to Energy-Parity, but replaces the Parity part by a MeanPayoff objective for a second reward dimension. I.e., one considers an MDP with 2-dimensional transition rewards, where the Energy condition applies to the first dimension and the MeanPayoff condition applies to the second dimension. (It can be generalized to higher dimensions d, where the MeanPayoff condition applies to all dimensions $2, 3, \ldots, d$.) This might look like a direct generalization of the Energy-Parity objective, since Parity games are reducible to MeanPayoff games [28, 25]. However, this reduction does not work in the context of these combined objectives when one considers stochastic systems like MDPs; see below. Non-stochastic Energy-MeanPayoff games have been studied in [9].

A sightly different objective has been studied in [17] who consider MDPs with d-dimensional rewards, where $d = d_1 + d_2$. The objective requires a strictly positive MeanPayoff *surely* in the first d_1 dimensions, and *almost surely* in the remaining d_2 dimensions. This objective is strictly stronger than Energy-MeanPayoff. E.g., a MeanPayoff of zero in the first dimension may or may not satisfy the Energy objective, but it never satisfies the objective in [17].

The objective studied in [6] aims to maximize the expected MeanPayoff (rather than the probability of it being strictly positive) while satisfying the energy constraint. However, unlike in our work, the reward function has a single dimension (i.e., both criteria apply to the same value) and ε-optimal strategies can require infinite memory.

Our contribution. We consider the Energy-MeanPayoff objective in MDPs with d-dimensional rewards. The first dimension needs to satisfy the Energy condition (never drop below 0), while each other dimension needs to have a *strictly* positive MeanPayoff. We show that almost surely winning strategies for Energy-MeanPayoff require only *finite* memory. [1] This

[1] Our results do *not* carry over to Energy-MeanPayoff objectives with *non-strict* inequalities where one just requires a MeanPayoff ≥ 0 almost surely. This needs infinite memory even for the case of $d = 2$, i.e., one energy-dimension and one MeanPayoff-dimension. It suffices to modify the counterexample for Energy-co-Büchi from [26, Page 4] such that a visit to a state with unfavorable color incurs a reward of -1 in the MeanPayoff-dimension.

is in contrast to the Energy-Parity objective where almost surely winning strategies require infinite memory in general [26, Page 4] (even for the simpler Energy-co-Büchi objectives). This also shows that Energy-Parity is not reducible to Energy-MeanPayoff in MDPs, unlike the reduction from Parity to MeanPayoff in [28, 25].

We show that almost surely winning strategies for Energy-MeanPayoff, if they exist, can be chosen as deterministic strategies with an exponential number of memory modes. The crucial property is that it suffices to remember the stored energy only up to some exponential upper bound. A small counterexample shows the corresponding exponential lower bound. Even for randomized strategies, an exponential number of memory modes is required, and this holds even for the case of small transition rewards in $\{-1, 0, +1\}$.

Although almost surely winning strategies are 'exponentially large' in this sense, their existence is still decidable in pseudo-polynomial time; cf. Section 7.

2 Preliminaries

A *probability distribution* over a countable set S is a function $f \colon S \to [0,1]$ with $\sum_{s \in S} f(s) = 1$. $\mathsf{supp}(f) \overset{\text{def}}{=} \{s \mid f(s) > 0\}$ denotes the support of f and $\mathcal{D}(S)$ is the set of all probability distributions over S. Given an alphabet Σ, let Σ^ω and Σ^* (Σ^+) denote the set of infinite and finite (non-empty) sequences over Σ, respectively. Elements of Σ^ω or Σ^* are called words.

MDPs and Markov chains. A *Markov Decision Process* (MDP) is a controlled stochastic directed graph $\mathcal{M} \overset{\text{def}}{=} (S, S_\Box, S_\bigcirc, E, P, \boldsymbol{r})$ where the set of vertices S (also called states) is partitioned into the states $S_\Box$ of the player $\Box$ (*Maximizer*), and chance vertices (aka random states) $S_\bigcirc$. Let $E \subseteq S \times S$ be the transition relation. We write $s{\longrightarrow}s'$ if $(s, s') \in E$ and assume that $\mathsf{Succ}(s) \overset{\text{def}}{=} \{s' \mid sEs'\} \neq \emptyset$ for every state s. The *probability function* P assigns each random state $s \in S_\bigcirc$ a distribution over its successor states, i.e., $P(s) \in \mathcal{D}(\mathsf{Succ}(s))$. We extend the domain of P to $S^* S_\bigcirc$ by $P(\rho s) \overset{\text{def}}{=} P(s)$ for all $\rho s \in S^+ S_\bigcirc$. A *Markov chain* is an MDP with only random states, i.e., $S_\Box = \emptyset$. In this paper we consider finite-state MDPs, i.e., the set of states S is finite.

Strategies. A *run* is an infinite sequence $s_0 s_1 \ldots \in S^\omega$ such that $s_i{\longrightarrow}s_{i+1}$ for all $i \geq 0$. A *path* is a finite prefix of a run. Let $\mathit{Runs}(\mathcal{M}) \overset{\text{def}}{=} \left\{\rho = (q_i)_{i \in \mathbb{N}} \,|q_i{\longrightarrow}q_{i+1}\right\}$ denote the set of all possible runs. A strategy of the player $\Box$ is a function $\sigma \colon S^* S_\Box \to \mathcal{D}(S)$ that assigns to every path $ws \in S^* S_\Box$ a probability distribution over the successors of s. If these distributions are always Dirac then the strategy is called *deterministic* (aka pure), otherwise it is called *randomized* (aka mixed). The set of all strategies of player $\Box$ in $\mathcal{M}$ is denoted by $\Sigma^{\mathcal{M}}$. A run/path $s_0 s_1 \ldots$ is compatible with a strategy σ if $s_{i+1} \in \mathsf{supp}(\sigma(s_0 \ldots s_i))$ whenever $s_i \in S_\Box$. Finite-memory strategies are a subclass of strategies using a finite set M of memory modes. A function $\mathsf{nxt} : \mathsf{M} \times S_\Box \mapsto \mathcal{D}(S)$ chooses a (distribution over) successor states based on the current memory mode and state and $\mathsf{upd} : \mathsf{M} \times E \mapsto \mathcal{D}(\mathsf{M})$ updates the memory mode upon observing a transition. Let $\sigma[\mathsf{m}]$ denote the finite-memory strategy σ starting in memory mode m. The set of all finite-memory strategies in $\mathcal{M}$ is denoted by $\Sigma_f^{\mathcal{M}}$. Strategies with memory $|\mathsf{M}| = 1$ are called *memoryless*. Memoryless deterministic (resp. randomized) strategies are called MD (resp. MR). By fixing some finite-memory strategy σ from some initial state in a finite-state MDP $\mathcal{M}$, we obtain a finite-state Markov chain, denoted by $\mathcal{M}^\sigma$.

Measure. An MDP $\mathcal{M}$ with initial state s_0 and strategy σ yields a probability space $(s_0 S^\omega, \mathcal{F}_{s_0}, \mathcal{P}^{\mathcal{M}}_{\sigma, s_0})$ where $\mathcal{F}_{s_0}$ is the σ-algebra generated by the cylinder sets $s_0 s_1 \ldots s_n S^\omega$ for

$n \geq 0$. The probability measure $\mathcal{P}^{\mathcal{M}}_{\sigma,s_0}$ is first defined on the cylinder sets. For $\rho = s_0 \dots s_n$, let $\mathcal{P}^{\mathcal{M}}_{\sigma,s_0}(\rho) \overset{\text{def}}{=} 0$ if ρ is not compatible with σ and otherwise $\mathcal{P}^{\mathcal{M}}_{\sigma,s_0}(\rho S^\omega) \overset{\text{def}}{=} \prod_{i=0}^{n-1} \tau(s_0 \dots s_i)(s_{i+1})$ where τ is σ or P depending on whether $s_i \in S_\Box$ or $S_\bigcirc$, respectively. If $\mathcal{M}$ is a Markov chain then there is only a single strategy, and we simply write $\mathcal{P}^{\mathcal{M}}_{s_0}$. By Carathéodory's extension theorem [4], this defines a unique probability measure on the σ-algebra. Given some reward function $v : s_0 S^\omega \to \mathbb{R}$, we write $\mathcal{E}(.)$ for the expectation w.r.t. $\mathcal{P}$ and v.

Objectives. General objectives are defined by real-valued measurable functions. However, we mostly consider indicator functions of measurable sets. Hence, our objectives can be described by measurable subsets $\mathtt{O} \subseteq S^\omega$ of runs starting at a given initial state. By $\mathcal{P}^{\mathcal{M}}_{\sigma,s}(\mathtt{O})$ we denote the payoff under σ, i.e., the probability that runs from s belong to $\mathtt{O}$. The value of a state is defined as $\mathtt{val}^{\mathcal{M}}_{\mathtt{O}}(s) \overset{\text{def}}{=} \sup_{\sigma \in \Sigma^{\mathcal{M}}} \mathcal{P}^{\mathcal{M}}_{\sigma,s}(\mathtt{O})$. For $\varepsilon > 0$ and state s, a strategy $\sigma \in \Sigma^{\mathcal{M}}$ is ε-optimal iff $\mathcal{P}^{\mathcal{M}}_{\sigma,s}(\mathtt{O}) \geq \mathtt{val}^{\mathcal{M}}_{\mathtt{O}}(s) - \varepsilon$. A 0-optimal strategy is called *optimal.* An MD/MR strategy is called *uniformly* ε-optimal (resp. uniformly optimal) if it is so from every start state. An optimal strategy from s is called *almost surely winning* if $\mathtt{val}^{\mathcal{M}}_{\mathtt{O}}(s) = 1$. By $\mathtt{AS}(\mathtt{O})$ (resp. $\mathtt{AS}_f(\mathtt{O})$) we denote the set of states that have an almost surely winning strategy (resp. an almost surely winning finite-memory strategy) for objective $\mathtt{O}$. For ease of presentation, we drop subscripts and superscripts wherever possible if they are clear from the context.

We use the syntax and semantics of the LTL operators [16] F (eventually), G (always) and X (next) to specify some conditions on runs. A reachability objective is defined by a set of target states $T \subseteq S$. A run $\rho = s_0 s_1 \dots$ belongs to $\mathsf{F}\, T$ iff $\exists i \in \mathbb{N}\, s_i \in T$. Similarly, ρ belongs to $\mathsf{F}^{\leq n} T$ (resp. $\mathsf{F}^{\geq n} T$) iff $\exists i \leq n$ (resp. $i \geq n$) such that $s_i \in T$. Dually, the safety objective $\mathsf{G}\, T$ consists of all runs which never leave T. We have $\mathsf{G}\, T = \neg \mathsf{F} \neg T$.

Energy/Reward/Counter-based objectives. Let $r : E \to \{-R, \dots, 0, \dots, R\}$ be a bounded function that assigns rewards to transitions. Depending on context, the sum of these rewards in a path can be viewed as energy, cost/profit or a counter. If $s {\longrightarrow} s'$ and $r((s, s')) = c$, we write $s \xrightarrow{c} s'$. Let $\rho = s_0 \xrightarrow{c_0} s_1 \xrightarrow{c_1} \dots$ be a run. We say that ρ satisfies

1. the k-*energy* objective $\mathsf{EN}(k)$ iff $\left(k + \sum_{i=0}^{n-1} c_i\right) \geq 0$ for all $n \geq 0$.
2. the l-*storage condition* $\mathtt{Infix}(l)$ if $l + \sum_{i=m}^{n-1} c_i \geq 0$ holds for every infix $s_m \xrightarrow{c_m} s_{m+1} \dots s_n$ of the run. Let $\mathsf{ST}(k, l)$ denote the set of runs that satisfy both the k-energy and the l-storage condition. Let $\mathsf{ST}(k) \overset{\text{def}}{=} \bigcup_l \mathsf{ST}(k, l)$. Clearly, $\mathsf{ST}(k) \subseteq \mathsf{EN}(k)$.
3. *Mean payoff* $\mathtt{MP}(\rhd c)$ for some constant $c \in \mathbb{R}$ iff $\left(\liminf_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} c_i\right) \rhd\ c$ for $\rhd \in \{<, \leq, =, \geq, >\}$.

A different way to consider the energy objective is to encode the energy level (the sum of the transition weights so far) into the state space and then consider the obtained infinite-state game with a safety objective.

An objective $\mathtt{O}$ is called *shift-invariant* iff for all finite paths ρ and plays $\rho' \in S^\omega$, we have $\rho\rho' \in \mathtt{O} \iff \rho' \in \mathtt{O}$. Mean payoff objectives are shift-invariant, but energy and storage/infix objectives are not.

Multidimensional reward-based objectives. Let $\mathbb{N}, \mathbb{Q}, \mathbb{R}$ denote the set of positive integers, rationals and reals respectively. For a d-dimensional real vector $\boldsymbol{\mu}$, let μ_i denote the i^{th} component of $\boldsymbol{\mu}$ for $1 \leq i \leq d$. Given two vectors $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{R}^d$, $\sim \in \{<, \leq, >, \geq, =\}$ we say $\boldsymbol{\mu} \sim \boldsymbol{\nu}$ if $\mu_i \sim \nu_i$ for every i. In particular, $\boldsymbol{\mu} > \mathbf{0}$ means that *every* component of $\boldsymbol{\mu}$ is strictly greater than 0. For a multidimensional reward function $\boldsymbol{r} : E \to [-R, R]^d$, we can consider any boolean combination of reward based objectives using any components of $\boldsymbol{r}$.

For instance, $\mathtt{O}_1 = \mathtt{EN}_1(k) \cap \mathtt{MP}_2(> 0)$ denotes the objective that contains all runs that satisfy $\mathtt{EN}(k)$ in the 1^{st} dimension and $\mathtt{MP}(> 0)$ in the 2^{nd} one. We denote conjunctions of the same objective across different dimensions in vectorized form, with the dimension information in the subscript. Therefore, $\mathtt{EN}_{[a,b]}(\boldsymbol{k}) \cap \mathtt{MP}_{[c,d]}(> \boldsymbol{x})$ denotes the runs where the $\mathtt{EN}_i(k_i)$ objective is satisfied for each $i \in [a, b]$ and the $\mathtt{MP}_j(> x_j)$ objective is satisfied for each $j \in [c, d]$. Given an infinite run $\rho = s_0 \xrightarrow{\boldsymbol{c_0}} s_1 \xrightarrow{\boldsymbol{c_1}} \ldots$, let $X_n(\rho) \overset{\text{def}}{=} s_n$ denote the n-th state. Let $\boldsymbol{Y_n}$ be the sum of the rewards in the first n steps, i.e., $\boldsymbol{Y_n}(\rho) \overset{\text{def}}{=} \sum_{i=0}^{n-1} \boldsymbol{c_i}$. These become random variables once an initial distribution and a strategy are fixed.

Size of an instance. Given an MDP $\mathcal{M} = (S, S_\Box, S_\bigcirc, E, P, \boldsymbol{r})$ with reward function $\boldsymbol{r} : E \to [-R, R]^d$, its size $|\mathcal{M}|$ is the number of bits used to describe it. Similarly for $|P|$. Transition probabilities and rewards can thus be stored in binary. We call a size pseudo-polynomial in $|\mathcal{M}|$ if it is polynomial for the case where R is 'small', i.e., if R is given in unary.

3 The Main Result

▶ **Theorem 1.** *Let $\mathcal{M} = (S, S_\Box, S_\bigcirc, E, P, \boldsymbol{r})$ be an MDP with d-dimensional rewards on the edges $\boldsymbol{r} : E \to [-R, R]^d$. For the multidimensional Energy-MeanPayoff objective $\mathtt{EN}_1(k) \cap \mathtt{MP}_{[2,d]}(> \boldsymbol{0})$ the following properties hold.*

1. *The existence of an almost-surely winning strategy implies the existence of an almost-surely winning finite-memory strategy.*
2. *Moreover, a deterministic strategy with an exponential number of memory modes is sufficient.*
3. *An exponential (in $|P|$) number of memory modes is necessary in general, even for randomized strategies, even for $|S| = 5$, $d = 2$ and $R = 1$.*

In the following three sections we prove items 1.,2.,3. of Theorem 1, respectively.

Here we sketch the main idea for the upper bound. Except in a special corner case where the energy fluctuates only in a bounded region, almost-surely winning strategies for Energy-MeanPayoff can be chosen among some particular strategies that alternate between two modes, playing two different memoryless strategies. This alternation keeps the balance between the Energy-part and the MeanPayoff-part of the objective. This is similar to almost-surely winning strategies for the Energy-Parity objective in [26]. In one mode, one plays a randomized memoryless strategy that almost surely yields a positive mean payoff in all dimensions (in case of Energy-Parity, instead of mean payoff it satisfies Parity almost surely). This is called the *Gain* phase. Whenever the energy level (the cumulative reward in dimension 1) gets dangerously close to zero, one switches to the other mode and plays a different memoryless strategy that focuses exclusively on getting the energy level up again, while temporarily neglecting the other part of the objective (Parity or Mean payoff, respectively). This is called a *Bailout*. Once the energy level is sufficiently high, one switches back to the Gain phase again. The crucial property is that, except in a null set, only finitely many Bailouts are required, and thus the temporary neglect of the second part of the objective does not matter in the long run. Such a strategy uses infinite memory, because it needs to remember the unbounded energy level. For Energy-Parity (and even Energy-co-Büchi) this cannot be avoided and finite-memory strategies do not work [26]. However, for Energy-MeanPayoff one can relax the requirements somewhat. Suppose that one records the stored energy only up to a certain bound b, i.e., one forgets about potential excess energy above b. In that case, one might have to do infinitely many Bailouts with high probability, most of which are unnecessary (but one does not know which ones). However, for a sufficiently large bound b, these superfluous Bailouts occur so infrequently that they do not compromise the MeanPayoff-part of the objective. The critical part of the proof is to show this property and an upper bound on b. Once this is established, one obtains a finite-memory strategy, because it suffices to record the energy level only in the range $[0, b]$ (plus one extra bit of memory to record the current phase, Gain or Bailout).

Note that the argument above is different from the one that justifies finite-memory ε-optimal strategies for *Energy-Parity* in [18]. These also record the energy only in a bounded region, but stop doing Bailouts after the upper bound has been visited. I.e., they do too few Bailouts, and thus incur an ε-chance of losing. In contrast, our almost-surely winning strategies for Energy-MeanPayoff rather do too many Bailouts, but sufficiently infrequently such that they don't compromise the objective.

4 Proof of Item 1

W.l.o.g, we assume that every state in $\mathcal{M}$ has an almost surely winning strategy for Energy-MeanPayoff for some initial energy level. (Otherwise, consider a suitably restricted sub-MDP.) For conciseness, we denote the objective by $\mathtt{O}(k) \overset{\text{def}}{=} \mathtt{EN}_1(k) \cap \mathtt{MP}_{[\mathbf{2},\boldsymbol{d}]}(\boldsymbol{>} \mathbf{0})$. Let

$$\mathtt{Win}(s) \overset{\text{def}}{=} \{k \mid s \in \mathtt{AS}(\mathtt{O}(k))\}, \quad i_s \overset{\text{def}}{=} \min(\mathtt{Win}(s))$$

denote the possible initial energy levels and the minimum initial energy level such that one can win almost surely from state s. In particular, i_s is well defined by our assumption on $\mathcal{M}$.

Towards a contradiction, assume that not all configurations are winnable with a finite-memory strategy. I.e., let $\mathtt{Win}_f(s) \overset{\text{def}}{=} \{k \mid s \in \mathtt{AS}_f(\mathtt{O}(k))\}$ denote the energy levels from which one can win almost surely with a *finite-memory* strategy from s, and assume that there is a state $s^\dagger$ such that $i_{s^\dagger} \notin \mathtt{Win}_f(s^\dagger)$. We then construct a finite-memory winning strategy from $s^\dagger$ for $\mathtt{O}(i_{s^\dagger})$, leading to a contradiction. Similar to i_s, let f_s denote the minimal k such that $k \in \mathtt{Win}_f(s)$ and ∞ if there is no such k.

▶ **Definition 2.** *We construct a new MDP $\mathcal{M}^*$ which abstracts away all the $\mathtt{Win}_f$ configurations. At every state s, the player gets the option to enter a winning sink state if the energy level is sufficiently large to win with finite memory, i.e., if the current energy level is at least f_s. The states of the MDP $\mathcal{M}^*$ will have two copies of each state s of $\mathcal{M}$, namely s and s'. Moreover, we add a new state s_{win}. All states s' are controlled by $\Box$ and every step $s_1 \longrightarrow s$ in the original MDP $\mathcal{M}$ is now mapped to a step $s_1 \longrightarrow s'$ with the same reward (and the same probability if s_1 was a random state). In s', the player has two choices: he can either go to s with reward $\mathbf{0}$ or go to s_{win} with reward $(-f_s, \mathbf{0})$. The latter choice is only available if $f_s < \infty$. s_{win} is a winning sink where $s_{win} \longrightarrow s_{win}$ with reward $\mathbf{1}$, i.e., reward $+1$ in all dimensions.*

The following lemma shows that the existence of almost surely winning (finite-memory) strategies coincides in $\mathcal{M}^*$ and $\mathcal{M}$.

▶ **Lemma 3.** *Let $s \in S$ and $k \in \mathbb{N}$, and consider the objective $\mathtt{O}(k)$. There exists an almost surely winning strategy σ^* from s in $\mathcal{M}^*$ if and only if there exists an almost surely winning strategy σ from s in $\mathcal{M}$. Moreover, if σ^* is finite-memory then σ can be chosen as finite-memory, and vice-versa.*

Proof. Towards the 'only if' direction, let σ^* be a strategy from s in $\mathcal{M}^*$ that is almost surely winning for $\mathtt{O}(k)$. We define a strategy σ from s in $\mathcal{M}$ that plays as follows. First σ imitates the moves of σ^* until (if ever) σ^* chooses a move $s_1' \to s_{\text{win}}$ with non-zero probability at some state s_1'. This is possible, since any finite path in $\mathcal{M}^*$ that does not contain s_{win} can be bijectively mapped to a path in $\mathcal{M}$. The only difference is that paths in $\mathcal{M}^*$ contain extra steps via primed states, which are skipped in the paths in $\mathcal{M}$. Moreover, the transition probabilities at random states coincide in $\mathcal{M}^*$ and $\mathcal{M}$. If σ^* chooses a move $s_1' \to s_{\text{win}}$ with non-zero probability at some state s_1' then the current energy level must be $\geq f_{s_1}$, because σ^* satisfies the energy objective almost surely (and thus even surely). Thus, in $\mathcal{M}$, there exists an almost surely winning finite-memory strategy $\hat{\sigma}$ for $\mathtt{O}(f_{s_1})$ from s_1. In this situation σ continues by playing $\hat{\sigma}$ from s_1. Therefore, σ satisfies the energy objective surely. Moreover, by shift invariance and the properties of $\hat{\sigma}$, it also satisfies the Mean payoff objective almost surely. Thus, σ satisfies $\mathtt{O}(k)$ almost surely. Finally, if σ^* is finite-memory then so is σ, because $\hat{\sigma}$ is also finite-memory.

Towards the 'if' direction, let σ be a strategy from s in $\mathcal{M}$ that is almost surely winning for $\mathtt{O}(k)$. We define a strategy σ^* from s in $\mathcal{M}^*$ that imitates the moves of σ. Moreover, at primed states q' it always goes to q (and never to s_{win}). Since the probabilities at random states coincide in $\mathcal{M}^*$ and $\mathcal{M}$, also the probabilities of the induced paths coincide. The only difference is that the runs in $\mathcal{M}^*$ contain extra steps via primed states and these extra steps carry reward zero. Thus, the mean payoff of a run in $\mathcal{M}^*$ is $1/2$ the mean payoff of the corresponding run in $\mathcal{M}$. However, this does not affect the property that the mean payoff is > 0 almost surely in either MDP. Thus, σ^* satisfies $\mathtt{O}(k)$ almost surely. Finally, if σ is finite-memory then so is σ^*. ◀

The next lemma shows that, in $\mathcal{M}^*$, it is impossible to satisfy Energy-MeanPayoff from s with arbitrarily high probability, unless one also allows arbitrarily large fluctuations in the energy level, or $f_s = i_s$. (Recall that f_s, i_s are defined relative to $\mathcal{M}$.)

▶ **Lemma 4.** *For every state s with $f_s > i_s$ and every $\ell \in \mathbb{N}$, there exists a $\delta_\ell > 0$ such that $\mathtt{val}^{\mathcal{M}^*}_{\mathtt{O}(i_s) \cap \mathtt{Infix}_1(\ell)}(s) \leq 1 - \delta_\ell$.*

Proof. Towards a contradiction, assume that $\mathtt{val}^{\mathcal{M}^*}_{\mathtt{O}(i_s) \cap \mathtt{Infix}_1(\ell)}(s) = 1$ for some ℓ.

$\mathtt{O}(i_s) \cap \mathtt{Infix}_1(\ell) = \mathtt{EN}_1(i_s) \cap \mathtt{MP}_{\boldsymbol{[2,d]}}(\boldsymbol{> 0}) \cap \mathtt{Infix}_1(\ell) = \mathtt{ST}_1(i_s, \ell) \cap \mathtt{MP}_{\boldsymbol{[2,d]}}(\boldsymbol{> 0})$. Therefore, we have $\mathtt{val}^{\mathcal{M}^*}_s\big(\mathtt{ST}_1(i_s, \ell) \cap \mathtt{MP}_{\boldsymbol{[2,d]}}(\boldsymbol{> 0})\big) = 1$. Below we prove that this objective has a finite-memory almost-surely winning strategy σ in $\mathcal{M}^*$. Consider a modified MDP $\mathcal{M}_1^*$ that encodes the energy level up to $i_s + \ell$ in the states. A step exceeding the upper energy bound $i_s + \ell$ results in a truncation to $i_s + \ell$, while a step leading to a negative energy leads to a losing sink. There exists a memoryless randomized (MR) strategy σ_1 in $\mathcal{M}_1^*$ from state (s, i_s) that wins $\mathtt{MP}_{\boldsymbol{[2,d]}}(\boldsymbol{> 0})$ almost surely, by Lemma 6. We can then carry σ_1 back to $\mathcal{M}^*$ as a finite-memory strategy σ with $i_s + \ell + 1$ memory modes such that $\mathcal{P}^{\mathcal{M}^*}_{\sigma,s}\big(\mathtt{ST}_1(i_s, \ell) \cap \mathtt{MP}_{\boldsymbol{[2,d]}}(\boldsymbol{> 0})\big) = 1$. By set inclusion, $\mathcal{P}^{\mathcal{M}^*}_{\sigma,s}(\mathtt{O}(i_s)) = 1$. By Lemma 3, there also exists a finite-memory strategy from s in $\mathcal{M}$ that is almost surely winning for $\mathtt{O}(i_s)$. This implies $f_s = i_s$, a contradiction to our assumption $f_s > i_s$. Hence, we obtain $\delta_\ell \stackrel{\text{def}}{=} 1 - \mathtt{val}^{\mathcal{M}^*}_{\mathtt{O}(i_s) \cap \mathtt{Infix}_1(\ell)}(s) > 0$. ◀

The following three lemmas show that almost surely winning strategies for Energy-MeanPayoff can be found by combining two different memoryless strategies for the simpler $\mathtt{Bailout}$ and $\mathtt{Gain}$ objectives.

First, we define the objective $\mathtt{Bailout}(k) \stackrel{\text{def}}{=} \mathtt{EN}_1(k) \cap \mathtt{MP}_1(> 0)$. Let $i_s^{\mathtt{Bailout}}$ denote the minimal energy value k with which one can almost surely satisfy $\mathtt{Bailout}(k)$ when starting from state s (or ∞ if it does not exist).

▶ **Lemma 5.** *[6, Lemma 3] Let $\mathcal{M}$ be an MDP. If $s \in \mathtt{AS}(\mathtt{Bailout}(k))$ for some $k \in \mathbb{N}$ then $i_s^{\mathtt{Bailout}} \leq 3 \cdot |\mathcal{M}| \cdot R$. Moreover, there exists a uniform MD strategy $\sigma^*_{\mathtt{Bailout}}$ which is almost surely winning $\mathtt{Bailout}(k)$ from every state $s \in \mathtt{AS}(\mathtt{Bailout}(k))$.*

We define the $\mathtt{Gain}$ objective as $\mathtt{MP}_{[1,d]}(> 0)$. The following lemma shows that an almost surely winning strategy $\sigma^*_{\mathtt{Gain}}$ for this objective can be chosen as memoryless randomized.

▶ **Lemma 6.** *[7, Proposition 5.1] There is a uniform MR strategy $\sigma^*_{\mathtt{Gain}}$ which is almost surely winning for $\mathtt{Gain}$ (or any subset of dimensions) from all states $s \in \mathtt{AS}(\mathtt{Gain})$.*

A difference between $\mathcal{M}^*$ and $\mathcal{M}$ is that if one can almost surely win Energy-MeanPayoff in $\mathcal{M}^*$ then one can also push the energy level arbitrarily high. This does not always hold in $\mathcal{M}$. (Consider, e.g., a single-state Markov chain with a single loop with reward 0 in the 1^{st} dimension and +1 in all other dimensions.) The difference comes from the loop at state s_{win} in $\mathcal{M}^*$ which has a strictly positive reward in all dimensions. Thus, the following lemma only holds for $\mathcal{M}^*$.

▶ **Lemma 7.** *In $\mathcal{M}^*$, there are two uniform memoryless strategies $\sigma^*_{\mathtt{Bailout}}$ and $\sigma^*_{\mathtt{Gain}}$ which, starting from any state $s \in \mathtt{AS}(\mathtt{O}(k))$, almost surely satisfy $\mathtt{Bailout}(k)$ and $\mathtt{Gain}$, respectively.*

Proof. Let $s \in \mathtt{AS}(\mathtt{O}(k))$. We show that $s \in \mathtt{AS}(\mathtt{Bailout}(k))$ and $s \in \mathtt{AS}(\mathtt{Gain})$. The existence of the memoryless strategies $\sigma^*_{\mathtt{Bailout}}$ and $\sigma^*_{\mathtt{Gain}}$ then follows from Lemma 5 and Lemma 6, respectively.

We assumed that all states s in $\mathcal{M}$ admit an almost surely winning strategy for Energy-MeanPayoff. By Lemma 3, this also holds for all states q in $\mathcal{M}^*$. Let $\sigma_q^\sharp$ denote an almost surely winning strategy from q for $\mathtt{O}(i_q)$ in $\mathcal{M}^*$ (without restrictions on memory).

Recall from Section 2 that the random variable X_t denotes the state at time t, and Y_t denotes the (d-dimensional) sum of the rewards until time t.

▷ Claim 8. For every state $q \in \mathcal{M}^*$ there exists some number of steps $n_q \in \mathbb{N}$ and a probability $p_q > 0$ such that

$$\mathcal{P}^{\mathcal{M}^*}_{\sigma_q^\sharp,q}\left(\bigcup_{j=0}^{n_q}((Y_j)_1 > i_{X_j} - i_q) \cup ((Y_j)_1 \geq f_{X_j} - i_q)\right) \geq p_q.$$

Proof. Towards a contradiction, assume that for all m

$$\mathcal{P}^{\mathcal{M}^*}_{\sigma_q^\sharp,q}\left(\bigcup_{j=0}^{m}((Y_j)_1 > i_{X_j} - i_q) \cup ((Y_j)_1 \geq f_{X_j} - i_q)\right) = 0.$$

Due to the second part of the union, this implies that never $(Y_j)_1 + i_q \geq f_{X_j}$. Since $\sigma_q^\sharp$ satisfies $\mathtt{EN}_1(i_q)$ almost surely, it can never choose the step to s_{win}. This implies $\mathcal{P}^{\mathcal{M}^*}_{\sigma_q^\sharp,q}(\mathsf{F} s_{\text{win}}) = 0$, i.e., X_j is always different from s_{win}. (The values f_s were initially defined with respect to states s of the original MDP $\mathcal{M}$, but the definition is naturally extended to the MDP $\mathcal{M}^*$, by giving the primed states the same value, i.e., $f_{s'} = f_s$. The state s_{win} does not appear in $\mathcal{M}$, but only in $\mathcal{M}^*$. We can extend the definition by having $f_{s_{\text{win}}} = 0$. However, this is not strictly required. The f_{X_j} is already defined, since X_j is always different from s_{win}.)

Since $\sigma_q^\sharp$ satisfies $\mathtt{EN}_1(i_q)$ almost surely, all runs always satisfy $(Y_j)_1 \geq i_{X_j} - i_q$ for all j. On the other hand, our assumption yields $\mathcal{P}^{\mathcal{M}^*}_{\sigma_q^\sharp,q}\left(\bigcup_{j=0}^{m}(Y_j)_1 > i_{X_j} - i_q\right) = 0$. This implies that $(Y_j)_1 = i_{X_j} - i_q$ for all j. Hence, in all runs the energy fluctuates by at most $\ell \overset{\text{def}}{=} 2\max_q i_q$.

Thus, $\mathcal{P}^{\mathcal{M}^*}_{\sigma^\sharp_q,q}(\mathtt{O}(i_q) \cap \mathtt{Infix}_1(\ell)) = 1$. Then Lemma 4 implies that $f_q = i_q$. Since $X_0 = q$ we have $f_{X_0} = f_q$ and thus $(Y_0)_1 \geq f_{X_0} - i_q = 0$. This contradicts our assumption, since the second part of the union is surely satisfied. $\lhd$

For any state q, let n_q, p_q denote the values from Claim 8.

Now we show that $s \in \mathtt{AS}(\mathtt{Bailout}(k))$. Define a strategy $\sigma_{\mathtt{Bailout}}$ which plays in phases, separated by resets. It remembers the number of steps $t \geq 0$ since last reset, the (under-approximated) sum of rewards Q_t and the current state X_t. The first phase starts at state s and $\sigma_{\mathtt{Bailout}}$ plays like $\sigma^\sharp_s$ until one of the following events occur.

1. There is enough energy such that it is safe to move to s_{win}, i.e., $(Q_t \geq f_{X_t} - i_s)$, or
2. The current energy level is strictly greater than the minimal required energy level of the current state, i.e., $(Q_t > i_{X_t} - i_s)$, or
3. n_s steps have elapsed, i.e., $(t = n_s)$.

If at any point Item 1 happens, then the strategy simply goes to s_{win}. If it is the case that Item 2 occurs before $t = n_s$, let's say at some time t', then the phase ends at t'. The sum of the rewards in the phase, between the last reset (where $t = 0$) and the current time is $\geq i_{X_{t'}} - i_s + 1$. If neither Item 1 nor Item 2 occurs before $t = n_s$, then the phase ends and we let $t' \overset{\text{def}}{=} t = n_s$. The sum of the rewards in this phase is then exactly $i_{X_{t'}} - i_s$. At the end of the phase $\sigma_{\mathtt{Bailout}}$ resets the number of steps $(t = 0)$, and Q_t to 0. In the following phase it moves according to $\sigma^\sharp_{X_{t'}}$ until the next reset.

$\sigma_{\mathtt{Bailout}}$ clearly satisfies $\mathtt{EN}_1(k)$ as it is a mix of energy safe strategies $\left(\sigma^\sharp_q\right)_{q \in S^*}$ and since we are starting from a safe energy level. By Claim 8, there is a positive probability (lower-bounded by $\min_q p_q > 0$) that either Item 1 or Item 2 happens in each phase.

Hence, unless event Item 1 occurs, Item 2 occurs infinitely often almost surely. Moreover, since the length of phases is upper bounded by $\max_q n_q$, it occurs frequently. We obtain $\mathcal{P}^{\mathcal{M}^*}_{\sigma_{\mathtt{Bailout}},s}\left(\mathtt{MP}_1 \geq \min_q\left(\frac{p_q}{n_q}\right) > 0 \,\middle|\, \neg\mathsf{F} s_{\text{win}}\right) = 1$. On the other hand, if s_{win} is reached, then $\mathtt{MP}_1$ holds by shift invariance and the definition of the positive rewards in the loop at s_{win}. Therefore, $\mathcal{P}^{\mathcal{M}^*}_{\sigma_{\mathtt{Bailout}},s}(\mathtt{EN}_1(i_s) \cap \mathtt{MP}_1(>0)) = 1$.

Now we show that $s \in \mathtt{AS}(\mathtt{Gain})$. We make use of the following strategies.

- $\sigma^\sharp_q$ which satisfies $\mathtt{EN}_1(k) \cap \mathtt{MP}_{[\mathbf{2},\boldsymbol{d}]}(> \mathbf{0})$ almost surely from q for every $k \geq i_q$.
- a uniform MD strategy $\sigma^*_{\mathtt{MP}_1}$ which satisfies $\mathtt{MP}_1(>0)$ almost surely from every state. It exists since $\mathtt{AS}(\mathtt{MP}_1(>0)) = S^*$ (where S^* is the set of states of $\mathcal{M}^*$), because $\mathcal{P}^{\mathcal{M}^*}_{\sigma_{\mathtt{Bailout}},s}(\mathtt{EN}_1(i_s) \cap \mathtt{MP}_1(>0)) = 1$.

From the former, we get probabilistic bounds on the achievable mean payoff in all the dimensions, *i.e.*, for all states s, and $0 \leq \varepsilon < 1$, there is a $d-1$ dimensional vector $\boldsymbol{\nu_\varepsilon} > \mathbf{0}$ such that $\mathcal{P}^{\mathcal{M}^*}_{\sigma^\sharp_s,s}\left(\mathtt{MP}_{[\mathbf{2},\boldsymbol{d}]} \geq \boldsymbol{\nu_\varepsilon}\right) \geq 1 - \frac{\varepsilon}{2}$. This follows from the fact that for any sequence of decreasing vectors $\boldsymbol{\nu_n} \to \mathbf{0}$ in $\mathbb{R}^{d-1}$, $\mathtt{MP}_{[\mathbf{2},\boldsymbol{d}]}(> \mathbf{0}) = \bigcup_n \mathtt{MP}_{[\mathbf{2},\boldsymbol{d}]}(\geq \boldsymbol{\nu_n})$ and continuity of measures. Furthermore, denoting by $\boldsymbol{Y_t}$ the sum of rewards in all dimensions until time t, there exists a sufficiently large bound $n_\varepsilon \in \mathbb{N}$ such that $\mathcal{P}^{\mathcal{M}^*}_{\sigma^\sharp_s,s}\left(\frac{(Y_t)_j}{t} \geq \frac{(\nu_\varepsilon)_j}{2}\right) \geq 1 - \varepsilon$ in each of the dimensions $j \in [2, d]$ for all $t \geq n_\varepsilon$ steps. This can be shown by observing that $\mathtt{MP}_j\left(\geq (\nu_\varepsilon)_j\right) = \bigcap_{k=1}^\infty \bigcup_{n=1}^\infty \bigcap_{t=n}^\infty \left(\frac{(Y_t)_j}{t} \geq (\nu_\varepsilon)_j \cdot \left(1 - \frac{1}{2^k}\right)\right)$ and using continuity of measures.

Similarly, there exists a bound $n^*_\varepsilon \in \mathbb{N}$ and value $\nu^*_\varepsilon > 0$ such that $\mathcal{P}_{\sigma^*_{\mathtt{MP}_1},s}\left(\frac{(Y_t)_1}{t} \geq \frac{\nu^*_\varepsilon}{2}\right) \geq 1 - \varepsilon$ after $t \geq n^*_\varepsilon$ steps for every state s.

Now consider the following strategy $\sigma_{\mathtt{Gain}}$, which switches between two phases.

Phase 1: If the current state is q, it moves according to $\sigma^\sharp_q$ for some number $\alpha > n_\varepsilon$ of steps. Then it switches to phase 2.

Phase 2: It moves according to $\sigma^*_{\texttt{MP}_1}$ for some number $\beta > n^*_\varepsilon$ of steps, and then switches back to phase 1.

The strategy $\sigma_{\texttt{Gain}}$ is a finite-memory strategy, since the lengths of the alternating phases are bounded by α and β, respectively. (Even if $\sigma^\sharp_q$ is an infinite-memory strategy, it can only use bounded memory in each phase.)

We fix $\sigma_{\texttt{Gain}}$ from the start state s and obtain a finite-state Markov chain. In every BSCC of this Markov chain, the expected mean payoff in the 1^{st} dimension will be

$$\geq \frac{-i^\sharp + \beta \cdot (1-\varepsilon) \cdot \left(\frac{\nu^*_\varepsilon}{2}\right) - \beta \cdot \varepsilon \cdot R}{\alpha + \beta}.$$

where $i^\sharp = \max_s i_s$ denotes the maximum (over all states) minimal safe energy.

Similarly, in every BSCC, the expected mean payoff in the j^{th} dimension for $j \geq 2$ can be lower-bounded by

$$\geq \frac{\alpha \cdot \left((1-\varepsilon) \cdot \left(\frac{(\nu_\varepsilon)_j}{2}\right) - \varepsilon \cdot R\right) - \beta \cdot R}{\alpha + \beta}.$$

By choosing ε sufficiently small, β sufficiently large to make the first term positive and $\alpha \gg \beta$ sufficiently large to make the second term positive, we can get positive expected mean payoff in all dimensions. Since this holds in every BSCC of the induced finite Markov chain, the objective `Gain` is satisfied almost surely. ◀

The following lemma shows the converse of Lemma 7. In $\mathcal{M}^*$, it is always possible to win $\texttt{O}(i_s)$ almost surely from s by playing a particular strategy $\sigma^*_{\texttt{alt},\texttt{Z}_\texttt{b},\texttt{Z}_\texttt{g}}$ which combines the two uniform memoryless strategies $\sigma^*_{\texttt{Bailout}}$ and $\sigma^*_{\texttt{Gain}}$. Let Z_b denote the minimal universally safe energy level for `Bailout`, i.e., $Z_b \overset{\text{def}}{=} \max_s \min\{k \mid s \in \texttt{AS}(\texttt{Bailout}(k))\}$. Moreover, let $Z_g > Z_b$ be a larger energy level at which our strategy switches from $\sigma^*_{\texttt{Bailout}}$ to $\sigma^*_{\texttt{Gain}}$.

Similarly to [26], we define an infinite-memory strategy $\sigma^*_{\texttt{alt},\texttt{Z}_\texttt{b},\texttt{Z}_\texttt{g}}$ that always records the current energy level and operates by switching between two phases. It starts by playing $\sigma^*_{\texttt{Gain}}$ (Gain-phase) if our starting energy level is sufficiently high ($\geq Z_b + R$), and otherwise starts by playing $\sigma^*_{\texttt{Bailout}}$ (Bailout-phase). In the `Bailout`-phase, the primary goal is to pump the energy level up until it is $\geq Z_g$, and then it switches to the `Gain`-phase. It enters the `Bailout`-phase again if the energy level drops below $Z_b + R$ (in which case it will still be $\geq Z_b$).

▶ **Lemma 9.** *There exists a $Z_g \in \mathbb{N}$ such that for every s in $\mathcal{M}^*$ the strategy $\sigma^*_{\texttt{alt},\texttt{Z}_\texttt{b},\texttt{Z}_\texttt{g}}$ is almost surely winning for $\texttt{O}(i_s)$ from s.*

Proof. The parameter Z_g is chosen sufficiently large such that there is a fixed non-zero probability that after every Bailout-phase one never needs another `Bailout`. (Thus, except in a null set there are only finitely many Bailouts.) The existence of such a finite Z_g is guaranteed by the fact that $\lim_{k\to\infty} \mathcal{P}_{\sigma^*_{\texttt{Gain}},s}(\texttt{O}(k)) = 1$. (Lemma 22). Eventually, except in a null set, $\sigma^*_{\texttt{alt},\texttt{Z}_\texttt{b},\texttt{Z}_\texttt{g}}$ plays `Gain` forever, thus satisfying $\texttt{O}(i_s)$ almost surely from s. ◀

Some combined objectives like Energy-Parity really require infinite memory for almost surely winning strategies [26]. However, we show that a sufficiently large *finite* memory is enough to win Energy-MeanPayoff almost surely. The idea is to modify the strategy $\sigma^*_{\texttt{alt},\texttt{Z}_\texttt{b},\texttt{Z}_\texttt{g}}$ such that it remembers the current energy only in the interval $[0, b]$, for some sufficiently large $b > Z_g$, and ignores any possible excess energy above b. This modified strategy is denoted by $\sigma^*_{\texttt{alt},\texttt{Z}_\texttt{b},\texttt{Z}_\texttt{g},\texttt{b}}$, and it has a finite set of memory modes $[0, b] \times \{0, 1\}$. The $\{0, 1\}$ part is used to remember the current phase ($\texttt{Gain} = 0$ or $\texttt{Bailout} = 1$). Then $\sigma^*_{\texttt{alt},\texttt{Z}_\texttt{b},\texttt{Z}_\texttt{g},\texttt{b}}[(u,x)]$ denotes the strategy $\sigma^*_{\texttt{alt},\texttt{Z}_\texttt{b},\texttt{Z}_\texttt{g},\texttt{b}}$ with current memory mode $(u,x) \in [0,b] \times \{0,1\}$.

The finite bound b on the remembered energy has the effect that $\sigma^*_{\texttt{alt},\texttt{Z}_\texttt{b},\texttt{Z}_\texttt{g},\texttt{b}}$ can no longer guarantee a fixed positive probability of not needing another Bailout after each Bailout-phase. Thus, one might have infinitely many Bailouts with positive probability. (Most of these are unnecessary, but one cannot be sure which ones). Unlike for Energy-Parity, where using infinitely many $\texttt{Bailout}$ phases can compromise the objective, the nature of the $\texttt{MP}_{[\mathbf{2},\boldsymbol{d}]}(\mathbf{> 0})$ objective allows us to use infinitely many Bailouts with non-zero probability, provided that they happen sufficiently infrequently.

By its construction, the strategy $\sigma^*_{\texttt{alt},\texttt{Z}_\texttt{b},\texttt{Z}_\texttt{g},\texttt{b}}[(i_s,x)]$ is energy-safe from every state s, every initial energy $\geq i_s$ and $x \in \{0,1\}$. It remains to show that it also satisfies $\texttt{MP}_{[\mathbf{2},\boldsymbol{d}]}(\mathbf{> 0})$ almost surely. Since $\sigma^*_{\texttt{alt},\texttt{Z}_\texttt{b},\texttt{Z}_\texttt{g},\texttt{b}}$ is finite-memory, it suffices to consider the induced finite Markov chain $\mathcal{A}$ and show that the expected mean payoff is strictly positive in every BSCC. I.e., we prove that $\mathcal{E}_{\sigma^*_{\texttt{alt},\texttt{Z}_\texttt{b},\texttt{Z}_\texttt{g},\texttt{b}},s}(\texttt{MP}_{[\mathbf{2},\boldsymbol{d}]}) > \mathbf{0}$ for a sufficiently large b. To this end, we consider the finite Markov chains $\mathcal{A}^{\texttt{Gain}}$ and $\mathcal{A}^{\texttt{Bailout}}$ obtained by fixing the memoryless strategies $\sigma^*_{\texttt{Gain}}$ and $\sigma^*_{\texttt{Bailout}}$ in $\mathcal{M}^*$, respectively. The application of $\sigma^*_{\texttt{alt},\texttt{Z}_\texttt{b},\texttt{Z}_\texttt{g},\texttt{b}}$ can then be seen as alternating between these two Markov chains based on hitting certain energy levels.

Let $T^{\texttt{Gain}}$ denote the random variable that measures the length of a Gain-phase, when starting at energy level Z_g and assuming that the energy it truncated at b. Similarly, $T^{\texttt{Bailout}}$ is the random variable that measures the length of a Bailout-phase when starting at energy level Z_b. (Here it does not matter that the energy is truncated at b, since the Bailout-phase ends when the energy reaches $Z_g < b$.) Since R can be > 1, the Bailout-phase might actually start at a slightly higher energy level $u \in [Z_b, Z_b+R-1]$, and thus $T^{\texttt{Bailout}}$ over-approximates the actual length of the Bailout-phase, which is conservative for our analysis. Similarly, the Gain phase might start with an energy slightly higher than Z_g, and $T^{\texttt{Gain}}$ under-approximates the length of the Gain-phase, which is again conservative. The random variables $(Y_{T^{\texttt{Gain}}})_i$ and $(Y_{T^{\texttt{Bailout}}})_i$ then measure the sum of the rewards the i^{th} dimension obtained during the Gain and Bailout phases, respectively.

The following lemma shows that the strategy $\sigma^*_{\texttt{alt},\texttt{Z}_\texttt{b},\texttt{Z}_\texttt{g},\texttt{b}}$ can attain a strictly positive mean payoff in all dimensions $i \in [2,d]$, provided that the expected reward during the Gain-phase is sufficiently large (positive) and the expected reward during the Bailout-phase (though possibly negative) is not too small.

▶ **Lemma 10.** *If there are constants $v_i^1 > 0$ and v_i^2 such that, for all $i \in [2,d]$ and states q*

$$\mathcal{E}^{\mathcal{M}^*}_{\sigma^*_{\texttt{alt},\texttt{Z}_\texttt{b},\texttt{Z}_\texttt{g},\texttt{b}}[(Z_g,0)],q}((Y_{T^{\texttt{Gain}}})_i) \geq v_i^1$$

$$\mathcal{E}^{\mathcal{M}^*}_{\sigma^*_{\texttt{alt},\texttt{Z}_\texttt{b},\texttt{Z}_\texttt{g},\texttt{b}}[(Z_b,1)],q}((Y_{T^{\texttt{Bailout}}})_i) \geq v_i^2$$

$$v_i^1 + v_i^2 > 0$$

then $\mathcal{E}^{\mathcal{M}^*}_{\sigma^*_{\texttt{alt},\texttt{Z}_\texttt{b},\texttt{Z}_\texttt{g},\texttt{b}}[\mathsf{m}],s}(\texttt{MP}_i) > 0$ for all s and $\mathsf{m} \in [i_s, b] \times \{0,1\}$.

Proof. By fixing the finite-memory strategy $\sigma^*_{\texttt{alt},\texttt{Z}_\texttt{b},\texttt{Z}_\texttt{g},\texttt{b}}$, we obtain a finite Markov chain. Consider any BSCC in this Markov chain. In this BSCC, except for a null set of runs, either no Bailouts happen or infinitely many. In the former case, this BSCC behaves like playing $\sigma^*_{\texttt{Gain}}$ forever, which attains a strictly positive mean payoff in all dimensions almost surely, and thus a strictly positive expected mean payoff in each dimension i. In the second case, almost surely there happen infinitely many Bailouts, each starting at an every level $\geq Z_b$. Then, by the finiteness of the BSCC, we obtain that $\mathcal{E}(T^{\texttt{Gain}}) < \infty$. Moreover, by the definition of $\sigma^*_{\texttt{Bailout}}$, the expected duration of the Bailout-phase is always finite, i.e., $\mathcal{E}(T^{\mathtt{Bailout}}) < \infty$. Thus, by linearity of expectations, $\mathcal{E}^{\mathcal{M}^*}_{\sigma^*_{\mathtt{alt},\mathtt{Z_b},\mathtt{Z_g},\mathtt{b}},s}(\mathtt{MP}_i) \geq (v_i^1 + v_i^2)/(\mathcal{E}(T^{\mathtt{Gain}}) + \mathcal{E}(T^{\mathtt{Bailout}})) > 0$. ◀

The following technical Lemma 11 (proof in Appendix B) shows that the constants v_i^1, v_i^2 from Lemma 10 exist. Recall that the finite Markov chains $\mathcal{A}^{\mathtt{Gain}}$ and $\mathcal{A}^{\mathtt{Bailout}}$ are obtained by fixing the memoryless strategies $\sigma^*_{\mathtt{Gain}}$ and $\sigma^*_{\mathtt{Bailout}}$ in $\mathcal{M}^*$, respectively. Let $x_{\min,1}$ and $x_{\min,2}$ denote the minimal occurring non-zero probabilities in these two Markov chains, respectively. (They come from solutions of linear programs and can be chosen as only exponentially small, i.e., described by a polynomial number of bits; cf. Appendix B). The proof works by applying general results about expected first passage times in truncated Markov chains to the induced Markov chains $\mathcal{A}^{\mathtt{Gain}}$ and $\mathcal{A}^{\mathtt{Bailout}}$. The general idea is that in the Gain-phase one has a general up drift in all dimensions, and in particular in the first (energy) dimension. It is thus unlikely to go down very far in the energy dimension, even if the energy is truncated at b. Thus, for a sufficiently large truncation point b (actually $b = Z_g + 1$ suffices), the expected time spent in the Gain-phase is very large relative to the expected time spent in the Bailout phase. More exactly, the former increases exponentially in b, while the latter is polynomial in b. For a sufficiently large b (exponential in $|\mathcal{M}^*|$), the condition $v_i^1 + v_i^2 > 0$ is met.

▶ **Lemma 11.** *Let $\mu_i > 0$ denote the lower bound on the mean payoff in dimensioni in any BSCC in the Markov chain $\mathcal{A}^{\mathtt{Gain}}$ with corresponding computable constants c_i, $g_{\mathtt{Gain}}$, $h_{\mathtt{Gain}}$ ((5), (9), (1)), and let μ denote the lower bound on the mean payoff in the 1st dimension in any BSCC of $\mathcal{A}^{\mathtt{Bailout}}$ with the corresponding constants $g_{\mathtt{Bailout}}, h_{\mathtt{Bailout}}$. All the above constants, except c_i, can be chosen as at most exponential in $|\mathcal{M}^*|$ and $1/(1-c_i) \in \mathcal{O}\big(\exp\big(\exp\big(|\mathcal{M}^*|^{\mathcal{O}(1)}\big)\big)\big)$.*

Then there are constants $0 < C_1 < 1$, $C_2 > 0$, $C_3 > 0$, $C_4 > 0$, $C_5 > 0$, all exponential in $|\mathcal{M}^*|$ and dependent only on $\mathcal{M}$, such that for $k \overset{def}{=} \frac{2\cdot|S^*|}{x_{\min,1}^{|S^*|}} \in \mathcal{O}\big(\exp\big(|\mathcal{M}^*|^{\mathcal{O}(1)}\big)\big)$, any $\delta \in (0,1)$ sufficiently small such that $(|S^*|+1)\cdot\big(\frac{1}{\delta}-1\big) + \big\lceil\big(\log_{c_i}(\delta\cdot(1-c_i))\big)\big\rceil \geq \frac{h_{\mathtt{Gain}}}{\mu_i}$ for all $2 \leq i \leq d$, one can choose
$Z_g \overset{def}{=} Z_b + R + k{\cdot}R + \max_i\big(R\cdot\lceil\log_{c_i}(\delta\cdot(1-c_i))\rceil - R + 1, h_{\mathtt{Gain}}\big) \in \mathcal{O}\Big(e^{|\mathcal{M}^*|^{\mathcal{O}(1)}}\cdot\log(1/\delta)\Big)$
and $b \overset{def}{=} Z_g + 1$ so that

$$\mathcal{E}^{\mathcal{M}^*}_{\sigma^*_{\mathtt{alt},\mathtt{Z_b},\mathtt{Z_g},\mathtt{b}}[(Z_g,0)],q}((Y_{T^{\mathtt{Gain}}})_i) \geq C_1\cdot\frac{1}{\delta} - C_2\log_2\left(\frac{1}{\delta}\right) - C_3 \qquad \overset{def}{=} v_i^1$$

$$\mathcal{E}^{\mathcal{M}^*}_{\sigma^*_{\mathtt{alt},\mathtt{Z_b},\mathtt{Z_g},\mathtt{b}}[(Z_b,1)],q}((Y_{T^{\mathtt{Bailout}}})_i) \geq -C_4\log_2\left(\frac{1}{\delta}\right) - C_5 \qquad \overset{def}{=} v_i^2$$

In particular, in order to satisfy the condition $v_i^1 + v_i^2 > 0$, it suffices to choose $1/\delta \in \mathcal{O}\big(\max(1/C_1, \max_{2\leq j\leq 5} C_j)^{\mathcal{O}(1)}\big)$. Since the constants C_j are exponential in $|\mathcal{M}^*|$, and by the conditions on the other constants above, the value Z_g, and hence the overall bound $b = Z_g + 1$, can be chosen such that $b \in \mathcal{O}\big(\exp\big(|\mathcal{M}^*|^{\mathcal{O}(1)}\big)\big)$.

Now we can prove the first item of our main result.

Proof of Theorem 1(Item 1). Towards a contradiction, we assume that there exists a state $s^\dagger$ such that there is no finite-memory almost surely winning strategy from $s^\dagger$ for $\mathtt{O}(i_{s^\dagger})$ in the MDP $\mathcal{M}$.

First we consider the MDP $\mathcal{M}^*$. The finite-memory strategy $\sigma^*_{\mathtt{alt},\mathtt{Z_b},\mathtt{Z_g},\mathtt{b}}[(i_{s^\dagger},1)]$ from $s^\dagger$ is energy-safe by construction and satisfies $\mathtt{EN}_1(i_{s^\dagger})$ surely. Now consider the finite Markov chain induced by fixing this finite-memory strategy in $\mathcal{M}^*$. By Lemma 10 and Lemma 11, for a sufficiently large (exponential) b it yields a strictly positive expected mean payoff $v_i^1 + v_i^2 > 0$

in every dimension $i \in [2,d]$ in every BSCC of this Markov chain. Since the Markov chain is finite, this implies that the mean payoff in every dimension $i \in [2,d]$ is strictly positive almost surely. Hence, $\mathcal{P}^{\mathcal{M}^*}_{\sigma^*_{\mathtt{alt},\mathtt{Z_b},\mathtt{Z_g},\mathtt{b}}[(i_{s^\dagger},1)],s^\dagger}\big(\mathtt{MP}_{[2,d]}(>0)\big) = 1$ and thus $\mathcal{P}^{\mathcal{M}^*}_{\sigma^*_{\mathtt{alt},\mathtt{Z_b},\mathtt{Z_g},\mathtt{b}}[(i_{s^\dagger},1)],s^\dagger}(\mathtt{O}(i_{s^\dagger})) = 1$. So there exists an almost surely winning finite-memory strategy from $s^\dagger$ for $\mathtt{O}(i_{s^\dagger})$ in $\mathcal{M}^*$. However, Lemma 3 then implies that there also exists an almost surely winning finite-memory strategy from $s^\dagger$ for $\mathtt{O}(i_{s^\dagger})$ in $\mathcal{M}$. Contradiction. ◀

▶ Remark 12. If $\sigma^*_{\mathtt{alt},\mathtt{Z_b},\mathtt{Z_g},\mathtt{b}}$ satisfies $\mathtt{O}(i_s)$ almost surely from some state s then it also satisfies the stronger objective $\mathtt{O}(i_s) \cap \mathtt{Infix}(b)$ almost surely. Consider a winning run induced by $\sigma^*_{\mathtt{alt},\mathtt{Z_b},\mathtt{Z_g},\mathtt{b}}$. While the true energy might sometimes be higher than b, the energy remembered by $\sigma^*_{\mathtt{alt},\mathtt{Z_b},\mathtt{Z_g},\mathtt{b}}$ is always $\le b$. Even with this conservative under-approximation of the energy, the run still satisfies the energy objective. Therefore, in any winning run induced by $\sigma^*_{\mathtt{alt},\mathtt{Z_b},\mathtt{Z_g},\mathtt{b}}$, the energy can never *decrease* by more than b. Thus, also $\mathtt{Infix}(b)$ is satisfied almost surely.

5 Proof of Item 2

Given some state s, let $\sigma = (\mathsf{M}, \mathsf{m}_0, \mathsf{upd}, \mathsf{nxt})$ be a finite-memory strategy that is almost surely winning for $\mathtt{O}(i_s)$ (which exists by Item 1). We show there exists an almost surely winning strategy σ' for $\mathtt{O}(i_s)$ such that the energy fluctuations are bounded by some constant which is exponential in $|\mathcal{M}|$.

First, inside any BSCC B of $\mathcal{M}^\sigma$, we construct an almost surely winning strategy σ_B and upper bound the minimal safe energy levels and energy fluctuation while following σ_B. Using this, we upper bound the energy fluctuations in paths before reaching a BSCC. We use the fact that the set of states and transitions that occur in any BSCC of a Markov chain induced by fixing some finite-memory strategy in an MDP is an end component of this MDP ([19, Theorem 3.2]).

▶ **Lemma 13.** *Let B be a BSCC of $\mathcal{M}^\sigma$ and let $\mathcal{M}(B)$ be the corresponding end component in $\mathcal{M}$ with states S_B and transitions E_B. Then there is a strategy σ_B, a bound $b_B \in \mathcal{O}\big(\exp\big(|\mathcal{M}(B)|^{\mathcal{O}(1)}\big)\big)$ such that for any state $q \in S_B$, there is a minimal safe energy level $j_q \overset{def}{=} i_q^{\mathcal{M}(B)} \le 3 \cdot |S_B| \cdot R$ such that $\mathcal{P}^{\mathcal{M}(B)}_{\sigma_B,q}(\mathtt{O}(j_q) \cap \mathtt{Infix}(b_B)) = 1$.*

Proof Sketch. (Full proof in Appendix C.). The idea is that for $\mathcal{M}(B)$ there are two cases. In the first case it behaves similar to $\mathcal{M}^*$ from Section 4, in the sense that it is possible to win $\mathtt{Gain}$ and $\mathtt{Bailout}$ almost surely, and thus Energy-MeanPayoff can be won almost surely by switching between the two strategies for $\mathtt{Gain}$ and $\mathtt{Bailout}$ like in the strategy $\sigma^*_{\mathtt{alt},\mathtt{Z_b},\mathtt{Z_g}}$. Then one can invoke Lemma 11 and Remark 12 on $\mathcal{M}(B)$ to get an exponential bound b_B such that $\mathcal{P}^{\mathcal{M}(B)}_{\sigma_B,q}(\mathtt{O}(j_q) \cap \mathtt{Infix}(b_B)) = 1$.

If the first case does not hold then $\mathcal{M}(B)$ is very restrictive, and one can show that the energy level fluctuations are bounded by a constant in $\mathcal{O}(|S_B| \cdot R)$. ◀

Since the minimal safe energy levels inside these end components are not too large, one can then bound the energy fluctuations in paths before they reach any such end component $\mathcal{M}(B)$. The following lemma is shown in Appendix C.

▶ **Lemma 14.** *Let T denote the union of all S_B of every BSCC B of $\mathcal{M}^\sigma$, as in Lemma 13. Then one can almost surely reach any state in T with the corresponding minimal safe energy level with energy fluctuations of at most $5 \cdot |S| \cdot R$.*

Proof of Theorem 1(Item 2). By Lemmas 13 and 14, for each state s, one can choose a strategy σ and some constant $b \in \mathcal{O}\big(\exp\big(|\mathcal{M}|^{\mathcal{O}(1)}\big)\big)$ such that $\mathcal{P}^{\mathcal{M}}_{\sigma,s}(\mathtt{O}(i_s) \cap \mathtt{Infix}(b)) = 1$.

This means if one encodes the energy levels between $[0, b]$ into the state space by discarding any excess energy above b and redirecting all the transitions which result in a negative energy to a losing sink (for $\mathtt{MP}_{[2,d]}(>0)$) and constructs this larger MDP $\mathcal{M}[0, b]$, then there is a strategy σ' such that $\mathcal{P}^{\mathcal{M}[0,b]}_{\sigma',(s,k)}\big(\mathtt{MP}_{[2,d]}(>0)\big) = 1$ for every $k \in [i_s, b]$. Then, by Lemma 6, there also exists a memoryless (MR) strategy σ^* in $\mathcal{M}[0, b]$ which is almost surely winning $\mathtt{MP}_{[2,d]}(>0)$ from (s, k).

We can carry the memoryless strategy σ^* in $\mathcal{M}[0, b]$ back to $\mathcal{M}$ as a finite-memory strategy $\sigma^*_{\mathcal{M}}$ with memory $[0, b]$. It stores the encoded under-approximated energy level from $\mathcal{M}[0, b]$ in its finite memory instead. Thus $\sigma^*_{\mathcal{M}}$ is a finite-memory strategy from s that satisfies $\mathtt{O}(i_s)$ almost surely, and the size of its memory is bounded by $b \in \mathcal{O}\big(\exp\big(|\mathcal{M}|^{\mathcal{O}(1)}\big)\big)$.

The strategy $\sigma^*_{\mathcal{M}}$ uses randomization, because σ^* from Lemma 6 is MR. However, the MR strategy σ^* for the mean payoff objective could be replaced by a deterministic strategy with an exponential number of memory modes. Hence the overall number of memory modes in the obtained deterministic version of $\sigma^*_{\mathcal{M}}$ is still only exponential. ◀

6 The Lower Bound (Proof of Item 3)

In the previous sections we have shown that finite memory suffices for almost surely winning strategies for the Energy-MeanPayoff objective. However, the required memory depends on the given MDP. We show that no fixed finite amount of memory is sufficient for all MDPs. In fact, the required memory is exponential in the transition probabilities even for an otherwise fixed 5-state MDP with just one controlled state, $R = 1$ and $d = 2$.

▶ **Definition 15.** *Let $1 > \delta > 0$ and $\mathcal{M}_\delta = (S, S_\Box, S_\bigcirc, E, P, \boldsymbol{r})$ be an MDP with 2-dimensional rewards. It has just one controlled state s with transitions $s \to s_l$ and $s \to s_r$. From s_l there are two transitions $e_1 = (s_l \to s_l^1)$ and $e_2 = (s_l \to s_l^2)$. Let $\mathcal{P}(e_1) = (1 + \delta)/2$ and $\mathcal{P}(e_2) = (1 - \delta)/2$ and $\boldsymbol{r}(e_1) = (+1, +1)$ and $\boldsymbol{r}(e_2) = (-1, -1)$. s_l^1 and s_l^2 are random states which each have just one transition back to s with probability 1 and reward $\mathbf{0}$. From s_r there is only one transition e_3 back to s with probability 1 and $\boldsymbol{r}(e_3) = (+1, -1)$.*

The following lemma directly implies the exponential lower bound on the number of memory modes in Theorem 1(Item 3).

▶ **Lemma 16.** *Consider the Energy-MeanPayoff objective. For every finite bound $m \in \mathbb{N}$ on the number of memory modes there exists a $\delta \overset{def}{=} 1/(6m) > 0$ such that the finite MDP $\mathcal{M}_\delta = (S, S_\Box, S_\bigcirc, E, P, \boldsymbol{r})$ from Definition 15 satisfies the following properties.*

1. *$\exists\sigma'\ \mathcal{P}^{\mathcal{M}_\delta}_{\sigma',s}(\mathtt{EN}_1(0) \cap \mathtt{MP}_{\mathbf{2}}(>0)) = 1$, i.e., it is possible to win almost surely from s in $\mathcal{M}_\delta$, even with initial energy 0.*
2. *For every finite-memory strategy σ with $\leq m$ memory modes we have $\mathcal{P}^{\mathcal{M}_\delta}_{\sigma,s}(\mathtt{EN}_1(k) \cap \mathtt{MP}_{\mathbf{2}}(>0)) = 0$ for every $k \in \mathbb{N}$, i.e., σ attains nothing in $\mathcal{M}_\delta$, regardless of the initial energy k.*
3. *For $\mathcal{M}_\delta$ we have $|S| = 5$, $d = 1$ and $R = 1$. The number of memory modes required for an almost-surely winning strategy in $\mathcal{M}_\delta$ is exponential in $|P|$ (and in $|\mathcal{M}_\delta|$).*

Proof. Towards item 1, consider a strategy σ' that plays as follows. It keeps a counter that records the current energy, which is initially 0. Whenever the current energy is 0, it plays $s \to s_r$, otherwise it plays $s \to s_l$. Thus σ' satisfies $\mathtt{EN}_1(0)$ surely from s. Since $\delta > 0$ it follows from the classic Gambler's ruin problem (with strictly positive expected gain, here in the first reward dimension) that σ' plays $s \to s_r$ only finitely often, except in a nullset of the runs.Therefore, the expected mean payoff (in the second dimension) under σ'

is $(1+\delta)/2 - (1-\delta)/2 = \delta > 0$. Hence $\mathcal{P}^{\mathcal{M}_\delta}_{\sigma',s}(\mathtt{MP}_{\mathbf{2}}(>0)) = 1$. Since the energy objective is satisfied surely, we obtain $\mathcal{P}^{\mathcal{M}_\delta}_{\sigma',s}(\mathtt{EN}_1(0) \cap \mathtt{MP}_{\mathbf{2}}(>0)) = 1$.

Towards item 2, let $\delta \overset{\text{def}}{=} 1/(6m) > 0$ and let σ be a finite-memory strategy with $\leq m$ memory modes. Consider the finite-state Markov chain $\mathcal{C}$ that is induced by playing σ from s in $\mathcal{M}_\delta$. This Markov chain has $\leq 5m$ states, since $\mathcal{M}$ has 5 states and σ has $\leq m$ memory modes. Let B be any BSCC of $\mathcal{C}$ that is reachable from s and the initial memory mode of σ. In particular, $|B| \leq 5m$. In B there must not exist any loop that does not contain s_r, because otherwise the energy objective cannot be satisfied almost surely. Thus every path in B of length $\geq 5m$ must contain s_r (and hence a reward $(+1,-1)$) at least once. Therefore, the expected mean payoff in B (in the second reward dimension) is $\leq 5m\delta - 1 = -1/6 < 0$. Since this holds in every reachable BSCC, we obtain $\mathcal{P}^{\mathcal{M}_\delta}_{\sigma,s}(\mathtt{MP}_{\mathbf{2}}(>0)) = 0$ and thus $\mathcal{P}^{\mathcal{M}_\delta}_{\sigma,s}(\mathtt{EN}_1(k) \cap \mathtt{MP}_{\mathbf{2}}(>0)) = 0$.

Towards item 3, the size of $\mathcal{M}_\delta$ follows from Definition 15. By items 1 and 2, the required number of memory modes m for an almost-surely winning strategy satisfies $m > 1/(6\delta)$. Since $|P| = \Theta(\log(1/\delta))$ and $|\mathcal{M}_\delta| = \Theta(|P|)$, we obtain $m = \Omega(\exp(|P|))$ and $m = \Omega(\exp(|\mathcal{M}_\delta|))$. ◀

The exponential lower bound on the required memory does not require probabilities encoded in binary like in Lemma 16. One can construct an equivalent example with polynomially many states where all transition probabilities are $1/2$. This is because one can encode exponentially small probabilities 2^{-k} with a chain of k extra states and transition probabilities $1/2$.

7 Computational Complexity

We have shown that the existence of an almost surely winning strategy for the Energy-MeanPayoff objective for a given state and initial energy level in an MDP implies the existence of a deterministic such strategy with exponentially many memory modes (unlike for Energy-Parity which requires infinite memory in general [26]).

A related problem is the decidability of the question whether a given state in an MDP and a given initial energy level admit an almost surely winning strategy for Energy-MeanPayoff. This problem is decidable in *pseudo-polynomial* time, using an algorithm very similar to the one for Energy-Parity presented in [26]. I.e., the time is polynomial, provided that the bound R on the rewards is given in unary. Transition probabilities in the MDP can still be represented in binary. The crucial point is that it suffices to witness the mere *existence* of an almost surely winning strategy, regardless of its memory. Basically, it suffices that the algorithm proves that the infinite-memory strategy $\sigma^*_{\mathtt{alt},\mathtt{Z_b},\mathtt{Z_g}}$ wins almost surely (plus a small extra argument about a corner case where the energy fluctuates only in a bounded region). The algorithm does not need to compute the bound b or to explicitly construct the finite-memory strategy $\sigma^*_{\mathtt{alt},\mathtt{Z_b},\mathtt{Z_g},\mathtt{b}}$.

▶ **Proposition 17.** *Let $\mathcal{M} = (S, S_\Box, S_\bigcirc, E, P, \boldsymbol{r})$ be an MDP with d-dimensional rewards on the edges $\boldsymbol{r} : E \to [-R, R]^d$. For any state s and $k \in \mathbb{N}$, the existence of an almost surely winning strategy from s for the multidimensional Energy-MeanPayoff objective $\mathtt{EN}_1(k) \cap \mathtt{MP}_{[\mathbf{2},\boldsymbol{d}]}(\mathbf{>0})$ is decidable in pseudo-polynomial time (i.e., polynomial for R in unary).*

Proof. The proof is similar to the one for Energy-Parity presented in [26]. We outline the differences below. First, in the corner case where it is impossible to pump the energy up arbitrarily high almost surely from some state q, the only possible way to win Energy-MeanPayoff (resp. Energy-Parity) almost surely (if at all) is by using a non-nullset of runs where the energy only ever fluctuates in a bounded region. In that case, the size of the energy fluctuations in these runs can safely be restricted to a region that is polynomial in $|S| \cdot R$, and thus pseudo-polynomial in $|\mathcal{M}|$ [26]. It thus suffices to win multi-dimensional $\mathtt{MP}_{[2,d]}(>0)$ almost surely in a derived MDP $\mathcal{M}'$ where the bounded energy is encoded into the states. Deciding this requires time polynomial in $|\mathcal{M}'|$ [13, 22] and thus pseudo-polynomial in $|\mathcal{M}|$. The winning situations of the corner case can then be encoded into $\mathcal{M}$, yielding a derived MDP $\mathcal{M}'$ of pseudo-polynomial size, where Energy-MeanPayoff can be won almost surely if and only if it can be won almost surely by a combination of Gain and Bailout strategies, i.e., by strategy $\sigma^*_{\mathtt{alt},\mathtt{Z_b},\mathtt{Z_g}}$. Therefore it suffices to compute the states (and minimal initial energy levels k) where Gain and Bailout(k) can be won almost surely. The objective $\mathtt{Bailout}(k) \overset{\text{def}}{=} \mathtt{EN}_1(k) \cap \mathtt{MP}_1(>0)$ is exactly the same as the Bailout objective analyzed in [26], and winning it almost surely is decidable in pseudo-polynomial time. Our objective $\mathtt{Gain} \overset{\text{def}}{=} \mathtt{MP}_{[1,d]}(>0)$ differs from the Gain objective considered in [26] (which was $\mathtt{MP}_1(>0) \cap \mathtt{Parity}$), but winning it almost surely is still decidable in polynomial time [13, 22] by solving a linear program. So overall the algorithm runs in pseudo-polynomial time. ◀

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.

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 Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

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 T. Brázdil, A. Kučera, and P. Novotný. Optimizing the expected mean payoff in energy Markov decision processes. In *International Symposium on Automated Technology for Verification and Analysis (ATVA)*, volume 9938 of *LNCS*, pages 32–49, 2016.

7 Tomáš Brázdil, Václav Brožek, Krishnendu Chatterjee, Vojtěch Forejt, and Antonín Kučera. Markov decision processes with multiple long-run average objectives. *Logical Methods in Computer Science*, 10, 2014.

8 Tomás Brázdil, Stefan Kiefer, and Antonín Kučera. Efficient analysis of probabilistic programs with an unbounded counter. *Journal of the ACM*, 61(6):41:1–41:35, 2014.

9 Véronique Bruyère, Quentin Hautem, Mickael Randour, and Jean-François Raskin. Energy Mean-Payoff Games. In Wan Fokkink and Rob van Glabbeek, editors, *30th International Conference on Concurrency Theory (CONCUR 2019)*, volume 140 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 21:1–21:17, Dagstuhl, Germany, 2019. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/LIPIcs.CONCUR.2019.21.

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

11 Arindam Chakrabarti, Luca De Alfaro, Thomas A. Henzinger, and Mariëlle Stoelinga. Resource interfaces. In *International Workshop on Embedded Software*, pages 117–133, 2003.

12 K. Chatterjee and T. Henzinger. A survey of stochastic ω-regular games. *Journal of Computer and System Sciences*, 78(2):394–413, 2012.

13 Krishnendu Chatterjee and Laurent Doyen. Energy and mean-payoff parity Markov decision processes. In *International Symposium on Mathematical Foundations of Computer Science (MFCS)*, volume 6907, pages 206–218, 2011.

14 Krishnendu Chatterjee, Thomas A. Henzinger, and Marcin Jurdziński. Mean-payoff parity games. In *Logic in Computer Science (LICS)*, pages 178–187, 2005.

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

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

17 Lorenzo Clemente and Jean-Francois Raskin. Multidimensional beyond worst-case and almost-sure problems for mean-payoff objectives. In *Proceedings of the 2015 30th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, page 257–268, 2015.

18 Mohan Dantam and Richard Mayr. Approximating the value of energy-parity objectives in simple stochastic games. In *48th International Symposium on Mathematical Foundations of Computer Science (MFCS 2023)*, pages 38:1–38:15, 2023. doi:10.4230/LIPIcs.MFCS.2023.38.

19 Luca De Alfaro. *Formal verification of probabilistic systems*. PhD thesis, Stanford University, 1997.

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

21 Dean Gillette. Stochastic games with zero stop probabilities. *Contributions to the Theory of Games*, 3:179–187, 1957.

22 Hugo Gimbert, Youssouf Oualhadj, and Soumya Paul. Computing optimal strategies for Markov decision processes with parity and positive-average conditions. Working paper or preprint, 2011.

23 Michael X. Goemans. An introduction to linear programming, 1994. URL: https://www.cs.cmu.edu/afs/cs/user/glmiller/public/Scientific-Computing/F-11/RelatedWork/Goemans-LP-notes.pdf.

24 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.

25 M. Jurdziński. Deciding the winner in parity games is in UP $\cap$ co-UP. *Information Processing Letters*, 68:119–124, 1998.

26 Richard Mayr, Sven Schewe, Patrick Totzke, and Dominik Wojtczak. MDPs with Energy-Parity Objectives. In *Logic in Computer Science (LICS)*. IEEE, 2017.

27 Richard Mayr, Sven Schewe, Patrick Totzke, and Dominik Wojtczak. Simple stochastic games with almost-sure energy-parity objectives are in NP and coNP. In *Proc. of Fossacs*, volume 12650 of *LNCS*, 2021. Extended version on arXiv. URL: https://arxiv.org/abs/2101.06989.

28 A. Puri. *Theory of hybrid systems and discrete event structures*. PhD thesis, University of California, Berkeley, 1995.

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

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

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

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

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

A Bounds in Markov Chains

This section shows some generic results for Markov chains with transition rewards. We show bounds on the expected arrival time (aka first passage time) of situations when the total reward reaches particular levels, under the condition that the total reward is truncated to remain inside some interval $[a, b]$. E.g., the total reward might hit the upper limit b many times (and be truncated there) before arriving at the lower level a for the first time. These bounds are later applied to Markov chains obtained by fixing certain finite-memory strategies in MDPs, and they are used in the proof of Lemma 11.

Let $\mathcal{A}$ be a strongly connected Markov chain with state space S, one-step transition probability matrix P and stationary distribution $\pi > \mathbf{0}$. Consider a reward function on the edges $r : E \to [-R, R]$ (alternately it can be seen as a vector in $[-R, R]^E$)such that the average reward gained in the limit is positive, i.e., $\mu \overset{\text{def}}{=} \sum_{e=(s,s')\in E} f_e \cdot r(e) > 0$ where $f_e \overset{\text{def}}{=} \pi(s) \cdot P(s)(s')$ denotes the long term relative frequency of the edge e. We begin by defining some functions on S^ω. Let $\rho = q_0q_1 \ldots$ be a generic infinite word.

Recall that X_n denotes the state of a Markov chain at time n and Y_n is the sum of the rewards until time n.

Let $x_{\min}$ denote the minimum occurring probability in $\mathcal{A}$. The size of the Markov chain with the reward structure $|\mathcal{A}|$ is defined as the total number of bits required to represent each state, edge, probability and reward in binary. We assume that all the probabilities are rational and rewards integers. Let

$$h \overset{\text{def}}{=} \frac{2|S|R}{x_{\min}^{|S|}} \tag{1}$$

▶ **Lemma 18.** *[8, Theorem 3.4] Let $u_s \overset{def}{=} \sum_{s'\in\mathsf{Succ}(s)} P(s)(s') \cdot r((s, s'))$ be the expected reward gained after taking an edge from state s. There exists $\nu \in [0, h]^S$ such that*

$$u + P\nu = \nu + \mathbf{1}\mu$$

where $\mathbf{1}$ on the RHS is a vector of all 1's.

▶ **Fact 1.** *Let ν be the vector from Lemma 18 and s be the start state, i.e., $X_0 = s$. Then the sequence of random variables given by*

$$M_n^s \overset{def}{=} Y_n + \nu(X_n) - n\mu$$

is a martingale for all s.

Since the average mean payoff $\mu > 0$ is strictly positive, $\liminf_{n\to\infty} Y_n = \infty$ almost surely. In the above setting, suppose that we bound the total reward gained to lie in some interval $[a, b]$ for some integers $a < 0 < b$, i.e., we define a new sequence of functions inductively as follows.

$$Y_0^{[a,b]} \overset{\text{def}}{=} Y_0 = 0$$
$$Y_n^{[a,b]} \overset{\text{def}}{=} \max(a, \min(b, Y_{n-1}^{[a,b]} + r((X_{n-1}, X_n)))) \quad \text{for } n \geq 1$$

Considering the sequence $Y_n^{[a,b]}$, let $T_a^{[a,b]}$, $T_b^{[a,b]}$ be the functions which denote the first hitting time of the left boundary a and right boundary b respectively. Clearly, $Y_n^{[a,b]} \leq Y_n$ before $T_a^{[a,b]}$, because the only possible difference is that $Y_n^{[a,b]}$ loses something when hitting the right border b. One of the advantages of $Y_n^{[a,b]}$ is that it can be described using only a finite number of bits for any n, because of its boundedness, whereas this is not the case for Y_n. This is useful in situations where such an under-approximation suffices instead of remembering the exact reward gained. However, the bounding of the reward also changes the behaviour of Y_n. For example, the probability of Y_n falling below a infinitely often is zero, i.e., $\mathcal{P}(\mathsf{GF}(Y_n \leq a)) = 0$ for a positive mean payoff $\mu > 0$. The same does not generally hold for $Y_n^{[a,b]}$, which might fall below a infinitely often almost surely.

We want to show that (on average) $Y_n^{[a,b]}$ hits the lower bound a much less frequently than the upper bound b. That is, we derive a lower bound on the expected time it takes to hit the lower bound a, and an upper bound on the expected time it takes to hit the upper bound b.

▶ Remark 19. Although we denote the functions by $T_a^{[a,b]}$, $T_b^{[a,b]}$ and $Y_n^{[a,b]}$ etc., note that there is an implicit assumption that the initial sum is 0 which lies between a and b. Therefore, the hitting times are actually parametrized by three numbers a, b and x such that $a < x < b$ given by $T_a^{[a,x,b]}$, $T_b^{[a,x,b]}$, $Y_n^{[a,x,b]}$. But we continue to represent with just the boundary points as all the functions are invariant under translation *i.e.*, , $T_a^{[a,x,b]} = T_{a'}^{[a',b']}$- where $a' = a - x$ and $b' = b - x$. Moreover, for $a < x_1 < x_2 < b$

$$T_a^{[a,x_1,b]} \leq T_a^{[a,x_2,b]}$$

Or in other form $T_{a-x_1}^{[a-x_1,b-x_1]} \leq T_{a-x_2}^{[a-x_2,b-x_2]}$ for all $a < x_1 < x_2 < b$.

A.1 Upper bound on the hitting time in the direction of drift

Given an initial state s, let the random variable T_b denote the first time Y_n is $\geq b > 0$.

$$T_b \stackrel{\text{def}}{=} \inf\{n \mid Y_n \geq b\} \tag{2}$$

Since the overall drift is in the positive direction, i.e., $\mu > 0$, it is immediate that the expectation of T_b is finite for every b. Also, the event $T_b = n$ can be determined by looking at the first n steps of any run, so T_b is a stopping time w.r.t the natural filtration of the Markov chain.

▶ **Fact 2.** *For all states s and $b > 0$, T_b is a stopping time and $\mathcal{E}_s(T_b) < \infty$.*

Now we show some bounds on this expected stopping time.

▶ **Lemma 20.** *For all states s and $b > 0$*

$$\frac{b-h}{\mu} \leq \mathcal{E}_s(T_b) \leq \frac{b+h+R}{\mu}$$

where h is the constant from Equation (1).

Proof. Since $\mathcal{E}_s(T_b) < \infty$ and the martingale M_n^s has bounded step size, we can apply optional stopping theorem to get

$$\mathcal{E}_s\big(M_{T_b}^s\big) = \mathcal{E}_s(M_0^s) = \nu(s)$$

Since $0 \leq \nu(s) \leq h$, it follows that

$$0 \leq \mathcal{E}_s(Y_{T_b} + \nu(X_{T_b}) - T_b\mu) \leq h.$$

Simplifying by using linearity of expectation and the fact that $b \leq Y_{T_b} < b + R$ and $0 \leq \nu(X_{T_b}) \leq h$, we get

$$\begin{aligned} 0 &\leq b + R + h - \mathcal{E}_s(T_b)\mu \\ b + 0 - \mathcal{E}_s(T_b)\mu &\leq h \end{aligned}$$

Rearranging terms and noting that $\mu > 0$ yields the required bounds. ◀

Lemma 20 shows that, starting from any state, the expected time to hit any upper total reward boundary $b > 0$ is asymptotically linear in b. To get the upper bound for $T_b^{[a,b]}$, observe that $\left(Y_n \leq Y_n^{[a,b]}\right) \mid (T_b \geq n)$ implying $T_b^{[a,b]} \leq T_b$. Hence we obtain the following as a corollary.

▶ **Corollary 21.**

$$\mathcal{E}_s\left(T_b^{[a,b]}\right) \leq \frac{b+h+R}{\mu}$$

A.2 Bounding the probability of going in the opposite direction of the drift

We now derive an upper bound on the probability that the total reward ever falls below some large negative number $a < 0$. This will be used later to derive the required inequalities for $T_a^{[a,b]}$.

Given $a < 0$, let the random variable T_a denote the first time that the total reward is $\leq a$.

$$T_a \overset{\text{def}}{=} \inf\{n \mid Y_n \leq a\} \tag{3}$$

Since the average reward $\mu > 0$ is positive, it is unlikely that Y_n ever hits large negative numbers. The following lemma quantifies this intuition. Recall that $h = \frac{2|S|R}{x_{\min}^{|S|}}$ and let

$$\eta \overset{\text{def}}{=} \mu + h + R \tag{4}$$

$$c \overset{\text{def}}{=} e^{\frac{-\mu^2}{2\eta^2}} \tag{5}$$

▶ **Lemma 22.** *For all $a \leq -h$ and states s we have*

$$\mathcal{P}_s^{\mathcal{A}}(T_a < \infty) \leq \frac{c^{\left\lceil \frac{|a|}{R} \right\rceil}}{1-c}.$$

Proof. Observe that the consecutive terms of the sequence M_n^s differ by at most η. Consider the event $T_a = n$. From our assumption $a \leq -h$ we obtain that $a + h \leq 0$, and thus

$$\begin{aligned} M_n^s - M_0^s &= (Y_n - Y_0) + (\nu(X_n) - \nu(X_0)) - n\mu \\ &\leq a + h - n\mu \\ &\leq -n\mu \end{aligned}$$

Hence, using the Azuma-Hoeffding inequality, we obtain

$$\mathcal{P}(T_a = n) \leq \mathcal{P}(M_n^s - M_0^s \leq -n\mu) \leq e^{\frac{-n^2\mu^2}{2n\eta^2}} = \left(e^{\frac{-\mu^2}{2\eta^2}}\right)^n$$

Since we have defined $c \overset{\text{def}}{=} e^{\frac{-\mu^2}{2\eta^2}}$, we see that $\mathcal{P}(T_a = n) \leq c^n$. Since $T_a \geq \left\lceil \frac{|a|}{R} \right\rceil$, we get that

$$\mathcal{P}(T_a < \infty) = \sum_{n=\left\lceil \frac{|a|}{R} \right\rceil}^{\infty} \mathcal{P}(T_a = n) \leq \frac{c^{\left\lceil \frac{|a|}{R} \right\rceil}}{1-c}.$$

◀

The above lemma provides a bound for the case of Y_n where the total reward is unrestricted. It is exponentially more unlikely that Y_n ever drops as low as a when $a \to -\infty$.

We need a lower bound on the expected time to hit the left bound a for the bounded random variable $Y_n^{[a,b]}$ (since we are interested in a finite-memory strategy). To do so, we first lower bound it by another variable which is simpler to analyse. Recall that $T_a^{[a,b]} = \inf\{n \mid Y_n^{[a,b]} \leq a\}$. Define a new sequence of random variables $Y_n'^{[a,b]}$ inductively as follows.

$$Y_0'^{[a,b]} = Y_0^{[a,b]} = 0$$

$$Y_n'^{[a,b]} = \begin{cases} Y_{n-1}'^{[a,b]} + r((X_{n-1}, X_n)) \text{ if } \; a < Y_{n-1}'^{[a,b]} + r((X_{n-1}, X_n)) < b & : \text{rule} \\ 0 \text{ if } \; Y_{n-1}'^{[a,b]} + r((X_{n-1}, X_n)) \geq b & : \text{reset} \\ a \text{ if } \; Y_{n-1}'^{[a,b]} + r((X_{n-1}, X_n)) \leq a & : \text{hit} \end{cases}$$

Intuitively, the behaviour of $Y_n'^{[a,b]}$ is similar to that of $Y_n^{[a,b]}$, except when it hits/exceeds b. Instead of clamping to b, $Y_n'^{[a,b]}$ 'resets' and behaves as if it is starting from the current state X_n. Let $T_a'^{[a,b]}$ denote the first time that $Y_n'^{[a,b]}$ hits the left bound a, i.e., $T_a'^{[a,b]} \overset{\text{def}}{=} \inf\{n \mid Y_n'^{[a,b]} \leq a\}$.

▷ Claim 23. For all $n \geq 0$ and $b > 0 > a$ we have $Y_n^{[a,b]} \geq Y_n'^{[a,b]}$. Consequently, $T_a^{[a,b]} \geq T_a'^{[a,b]}$.

Proof. By induction on n.

In the base case $n = 0$ we have $Y_0^{[a,b]} = Y_0'^{[a,b]}$.

For the induction step let $n > 0$.

If $Y_{n-1}'^{[a,b]} + r((X_{n-1}, X_n)) \geq b$ then $Y_{n-1}^{[a,b]} + r((X_{n-1}, X_n)) \geq b$, by induction hypothesis. Then $Y_n'^{[a,b]} = 0 < b = Y_n^{[a,b]}$.

Else if $Y_{n-1}'^{[a,b]} + r((X_{n-1}, X_n)) \leq a$, then $Y_n'^{[a,b]} = a \leq Y_n^{[a,b]}$.

Otherwise, we have $a < Y_n'^{[a,b]}) = Y_{n-1}'^{[a,b]} + r((X_{n-1}, X_n)) < b$ and by induction hypothesis $Y_n'^{[a,b]} = Y_{n-1}'^{[a,b]} + r((X_{n-1}, X_n)) \leq \min(b, Y_{n-1}^{[a,b]} + r((X_{n-1}, X_n))) = Y_n^{[a,b]}$. ◀

To lower bound $\mathcal{E}_s\left(T_a'^{[a,b]}\right)$, we split each run at the reset points (when b is hit or exceeded) and evaluate the expected number of resets that happen before hitting a. Formally, let $V_a'^{[a,b]}$ be the random variable denoting the number of resets before hitting a.

$$V_a'^{[a,b]} \overset{\text{def}}{=} \sum_{i=1}^{T_a'^{[a,b]}} \mathbb{1}_{\left(Y_{i-1}'^{[a,b]} + r((X_{i-1}, X_i)) \geq b\right)}.$$

We analogously define the random variables $V_a^{[a,b]}$ and V_a^b for the non-primed random variable $Y_n^{[a,b]}$ and the unbounded random variable Y_n, respectively.

For the argument to work, we have to space the bounds a and b sufficiently far apart. In the rest of the section we assume that $a \leq -h < 0 < b$.

Since the step size is bounded by R, the constants $\alpha \overset{\text{def}}{=} \lceil \frac{|a|}{R} \rceil$ and $\beta \overset{\text{def}}{=} \lceil \frac{|b|}{R} \rceil$ are universal lower bounds on the minimum time it takes to hit the left bound a and the right bound b, respectively.

$$\mathcal{E}_s\left(T_a'^{[a,b]}\right) = \sum_{n=0}^{\infty} \mathcal{P}_s\left(T_a'^{[a,b]} > n\right)$$

Since hitting a takes at least $\alpha > 0$ steps, the probability $\mathcal{P}_s\left(T_a'^{[a,b]} > n\right) = 1$ for all $0 \leq n \leq \alpha - 1$. Thus, the summation can be simplified to

$$\begin{aligned}
&\alpha + \sum_{n=0}^{\infty} \mathcal{P}_s\left(T_a'^{[a,b]} > n + \alpha\right) \\
=&\alpha + \sum_{j=0}^{\infty} \sum_{k=0}^{\beta-1} \mathcal{P}_s\left(T_a'^{[a,b]} > j \cdot \beta + k + \alpha\right) \\
\geq&\alpha + \sum_{j=0}^{\infty} \beta \cdot \mathcal{P}_s\left(T_a'^{[a,b]} > (j+1) \cdot \beta + \alpha - 1\right) \\
=&\alpha + \sum_{j=0}^{\infty} \beta \cdot \mathcal{P}_s\left(T_a'^{[a,b]} \geq (j+1) \cdot \beta + \alpha\right) \\
\geq&\alpha + \sum_{j=0}^{\infty} \beta \cdot \mathcal{P}_s\left(T_a'^{[a,b]} \geq (j+1) \cdot \beta + \alpha \wedge V_a'^{[a,b]} \geq j+1\right) \\
\geq&\alpha + \sum_{j=0}^{\infty} \beta \cdot \mathcal{P}_s\left(V_a'^{[a,b]} \geq j+1\right)
\end{aligned}$$

where the last inequality is justified by the fact that resetting at least $j+1$ times implies that the time taken to hit the left bound a would be at least $(j+1) \cdot \beta + \alpha$.

▷ Claim 24. Let $0 < \delta < 1$ and let c be as in Equation (5). By choosing $a \overset{\text{def}}{=} \min(-R\lceil(\log_c(\delta \cdot (1-c)))\rceil + R - 1, -h)$ we obtain $\mathcal{P}_s\left(V_a'^{[a,b]} = 0\right) \leq \delta$ for any start state s. Moreover, it holds that $\mathcal{P}_s\left(V_a'^{[a,b]} \geq j+1\right) \geq (1-\delta)^{j+1}$.

Proof. From our choice of a and Lemma 22, it follows that $\mathcal{P}_s^{\mathcal{A}}(T_a < \infty) \leq \frac{c^{\left\lceil \frac{|a|}{R} \right\rceil}}{1-c} \leq \delta$. Consider the event $V_a'^{[a,b]} = 0$ when starting from s. This means any run in this event doesn't hit the reset transitions, which implies that the probability of this event doesn't change when considering Y_n or $Y_n^{[a,b]}$ instead of the sequence $Y_n'^{[a,b]}$.

$$\mathcal{P}_s\left(V_a'^{[a,b]} = 0\right) = \mathcal{P}_s\left(V^{[a,b]} = 0\right) = \mathcal{P}_s\left(V_a^b = 0\right).$$

But the event $V_a^b = 0$ is exactly equivalent to the event $T_a < T_b$ in $\mathcal{A}$ which further implies that $T_a < \infty$ in $\mathcal{A}$. Therefore, we have that

$$\mathcal{P}_s\left(V'^{[a,b]} = 0\right) = \mathcal{P}_s(T_a < T_b) \leq \mathcal{P}_s^{\mathcal{A}}(T_a < \infty) \leq \delta. \tag{6}$$

We can then prove the required claim by induction on j.

Base case $j = 0$: $\mathcal{P}_s\Big(V_a'^{[a,b]} \geq 1\Big) = \mathcal{P}_s\Big(V_a'^{[a,b]} > 0\Big) = 1 - \mathcal{P}_s\Big(V_a'^{[a,b]} = 0\Big) \geq 1 - \delta$, by Equation (6).

Induction step:

$$\begin{aligned}\mathcal{P}_s\Big(V_a'^{[a,b]} \geq j+2\Big) &= \mathcal{P}_s\Big(V_a'^{[a,b]} \geq j+2, V_a'^{[a,b]} \geq j+1\Big)\\ &= \mathcal{P}_s\Big(V_a'^{[a,b]} \geq j+2 \,|\, V_a'^{[a,b]} \geq j+1\Big) \cdot \mathcal{P}_s\Big(V_a'^{[a,b]} \geq j+1\Big)\end{aligned}$$

Let $T_{b,j}'^{[a,b]}$ denote the time taken for the j^{th} visit to energy level b in $\mathcal{A}$. It is clear that $T_{b,j}'^{[a,b]}$ is a stopping time for every j. Using strong Markov property, one can simplify the above conditional probability to get the required result.

$$\begin{aligned}&\geq (1-\delta)^{j+1} \cdot \sum_{s' \in S} \mathcal{P}_s\Big(V_a'^{[a,b]} \geq j+2 \,|\, V_a'^{[a,b]} \geq j+1, X_{T_{b,j+1}'^{[a,b]}} = s'\Big) \cdot \mathcal{P}_s\Big(X_{T_{b,j+1}'^{[a,b]}} = s' \,|\, V_a'^{[a,b]} \geq j+1\Big)\\ &= (1-\delta)^{j+1} \cdot \sum_{s' \in S} \mathcal{P}_{s'}\Big(V_a'^{[a,b]} \geq 1\Big) \cdot \mathcal{P}_s\Big(X_{T_{b,j+1}'^{[a,b]}} = s' \,|\, V_a'^{[a,b]} \geq j+1\Big)\\ &\geq (1-\delta)^{j+1} \cdot \sum_{s' \in S} (1-\delta) \cdot \mathcal{P}_s\Big(X_{T_{b,j+1}'^{[a,b]}} = s' \,|\, V_a'^{[a,b]} \geq j+1\Big)\\ &= (1-\delta)^{j+2}\end{aligned}$$

◀

Thus, we get that

$$\begin{aligned}\mathcal{E}_s\Big(T_a^{[a,b]}\Big) \geq \mathcal{E}_s\Big(T_a'^{[a,b]}\Big) &\geq \alpha + \sum_{j=0}^{\infty} \beta \cdot \mathcal{P}_s\Big(V_a'^{[a,b]} \geq j+1\Big)\\ &\geq \alpha + \sum_{j=0}^{\infty} \beta \cdot (1-\delta)^{j+1}\\ &= \alpha + \beta \cdot \frac{1-\delta}{\delta}\\ &= \beta \cdot \frac{1}{\delta} + (\alpha - \beta). \end{aligned} \tag{7}$$

We aim to make the interval $[a, b]$ as small as possible (since later the size of the memory in our strategies will be proportional to $|b - a|$). Due to the different influences of the parameters a and b, it is better to fix b as 1 and make a smaller (more negative). The β will then be 1 as well.

▶ **Lemma 25.** *For $0 < \delta < 1$ and $a = \min(-R\lceil(\log_c(\delta \cdot (1-c)))\rceil + R - 1, -h)$ we have*

$$\mathcal{E}_s\Big(T_a^{[a,1]}\Big) \geq \frac{1}{\delta} + \lceil(\log_c(\delta \cdot (1-c)))\rceil - 1.$$

Proof. Since $b = 1$ and $|a| \geq R(\lceil(\log_c(\delta \cdot (1-c)))\rceil - 1) + 1$, $\beta = 1$ and $\alpha = \lceil\frac{|a|}{R}\rceil \geq \lceil(\log_c(\delta \cdot (1-c)))\rceil$. Substituting these values in Equation (7) gives the required bound. ◀

As $\delta \to 0$, the expected hitting time grows as $\approx \frac{1}{\delta} = \exp(\log(1/\delta))$, i.e., exponentially in $\log(\frac{1}{\delta})$, whereas the memory $\approx a$ required to achieve this lower bound would be proportional to $R\Big(\frac{\log(1/\delta)}{\log(1/c)}\Big) = \frac{2\eta^2 R}{\mu^2}\log(1/\delta)$, linear in $\log(1/\delta)$.

A.3 General Markov Chains

To get a lower bound on $\mathcal{E}_s\left(T_a^{[a,b]}\right)$ when $\mathcal{A}$ is not strongly connected, one has to account for the time spent in transient states. Fortunately, the probability to spend a large amount of time outside a BSCC falls exponentially and this allows the analysis from the previous subsections to carry over with a minimal increase in the size of the interval $[a, b]$ required for general Markov chains. Assume $\texttt{AS}(\texttt{MP}(>0)) = S$, i.e., every BSCC of $\mathcal{A}$ has positive average mean payoff. Let $C \subseteq S$ denote all the recurrent states. Compute the constants η_G, μ_G, c_G, h_G from Lemma 22 for each BSCC G and let η, μ, c, h be the maximum over all BSCC's. Let T_C denote the hitting time of some BSCC. For some positive integer k, let Z_k denote the event $T_C \leq k$. By the tower property, one has

$$\mathcal{E}_s\left(T_a^{[a,b]}\right) = \mathcal{E}_s\left(\mathcal{E}\left(T_a^{[a,b]} \mid 1_{Z_k}\right)\right) = \mathcal{E}_s\left(T_a^{[a,b]} \mid \overline{Z_k}\right) \cdot \mathcal{P}_s(\overline{Z_k}) + \mathcal{E}_s\left(T_a^{[a,b]} \mid Z_k\right) \cdot \mathcal{P}_s(Z_k).$$

Since we are only interested in a lower bound, we can ignore the low probability event $\overline{Z_k}$ as $T_a^{[a,b]}$ is a non-negative random variable.

$$\mathcal{E}_s\left(T_a^{[a,b]}\right) \geq \mathcal{E}_s\left(T_a^{[a,b]} \mid Z_k\right) \cdot \mathcal{P}_s(Z_k). \tag{8}$$

If $x_{\min}$ is the minimum occurring probability in $\mathcal{A}$, then let

$$g \overset{\text{def}}{=} \exp\left(\frac{-x_{\min}^{|S|}}{|S|}\right) \tag{9}$$

▶ **Lemma 26.** *Let $y_{\min}$ denote the minimum occurring probability in $\mathcal{A}$ outside every BSCC. Then there exists $0 \leq g < 1$ such that for all $k > |S|$,*

$$\mathcal{P}_s(\overline{Z_k}) \leq 2 \cdot g^k.$$

Proof. The proof is similar to [8, Lemma 5.1]. Assume $y_{\min} \neq 1$ (the other case is trivial $\mathcal{P}_s(\overline{Z_k}) = 0$). This implies $y_{\min} \leq \frac{1}{2}$. Let $n = |S|$. From any state s, there will be a path of length at most $n-1$ to a state in C, $\implies$ for all states s, $\mathcal{P}_s(T_C < n) \geq y_{\min}^{n-1} \geq y_{\min}^n$. Dividing the run into segments of length $n-1$, one gets

$$\begin{aligned}
\mathcal{P}_s(\overline{Z_k}) &= \mathcal{P}_s(T_C > k) \\
&\leq \mathcal{P}_s(T_C \geq k) \\
&\leq (1 - y_{\min}^n)^{\lfloor \frac{k-1}{n-1} \rfloor} \\
&\leq 2 \cdot \left(\exp\left(\frac{1}{n}\log(1 - y_{\min}^n)\right)\right)^k \\
&\leq 2 \cdot g^k \left(g = \exp\left(\frac{-y_{\min}^n}{n}\right)\right)
\end{aligned}$$

◀

From the above lemma, we get a lower bound on $\mathcal{P}_s(Z_k)$ for $k > n$. Before computing a lower bound on $\mathcal{E}_s\left(T_a^{[a,b]} \mid Z_k\right)$, we choose a to be sufficiently negative so that it is never the case that $T_a^{[a,b]} \leq k$.

▶ **Lemma 27.** *For any $0 < \delta < 1$, choosing $a = \min(-R\lceil(\log_c(\delta \cdot (1-c)))\rceil + R - 1, -h) - k \cdot R$ and $b = 1$*

$$\mathcal{E}_s\left(T_a^{[a,b]} \mid Z_k\right) \geq (k+1) \cdot \left(\frac{1}{\delta} - 1\right) + \lceil(\log_c(\delta \cdot (1-c)))\rceil$$

Proof. Since $T_C \leq k < T_a^{[a,b]} \mid Z_k,\, Y_{T_C}^{[a,b]} \geq -k \cdot R$.

Let $I \stackrel{\text{def}}{=} \{(y,q) \mid Y_{T_C}^{[a,b]} = y \wedge X_{T_C} = q \wedge \mathcal{P}_s\left(Y_{T_C}^{[a,b]} = y, X_{T_C} = q \mid Z_k\right) > 0\}$

$$\mathcal{E}_s\left(T_a^{[a,b]} \mid Z_k\right) = \sum_{(y,q)\in I} \mathcal{E}_s\left(T_a^{[a,b]} \mid Z_k,\, Y_{T_C}^{[a,b]} = y,\, X_{T_C} = q\right) \cdot \mathcal{P}_s\left(Y_{T_C}^{[a,b]} = y, X_{T_C} = q \mid Z_k\right)$$

Let $E(y,q)$ denote the event $Z_k \cap Y_{T_C}^{[a,b]} = y \cap X_{T_C} = q$

$$\begin{aligned}\mathcal{E}_s\left(T_a^{[a,b]} \mid E(y,q)\right) &= \mathcal{E}_s(T_C \mid E(y,q)) + \mathcal{E}_q\left(T_{a-g}^{[a-y,b-g]}\right)\\ &\geq \mathcal{E}_q\left(T_{a+k\cdot R}^{[a+k\cdot R,b+k\cdot R]}\right)\\ &\geq (k+1)\cdot\left(\frac{1}{\delta}-1\right) + \lceil(\log_c(\delta\cdot(1-c)))\rceil \quad ((7))\end{aligned}$$

Summing over all mutually exclusive events, one obtains the specified bound. ◀

Using Lemmas 26 and 27, one gets the following result.

▶ **Lemma 28.** *For any $k > n$, $0 < \delta < 1$, with $a = \min(-R\lceil(\log_c(\delta\cdot(1-c)))\rceil + R - 1, -h) - k\cdot R$ and $b = 1$*

$$\mathcal{E}_s\left(T_a^{[a,b]}\right) \geq (1 - 2\cdot g^k)\cdot\left((k+1)\cdot\left(\frac{1}{\delta}-1\right) + \lceil(\log_c(\delta\cdot(1-c)))\rceil\right)$$

$$\mathcal{E}_s\left(T_b^{[a,b]}\right) \leq |S| + \frac{2}{1-g} + \frac{|a|+b+h+R}{\mu}$$

where g and h are computable constants dependent only on $\mathcal{A}$ and c depends on $\mathcal{A}$ along with reward function r. Furthermore, if $r_2 : E \to \{-R,\ldots,0,\ldots,R\}$ is an additional reward function with positive mean payoff of at least μ_2 in every BSCC, then assuming δ is small enough such that $(|S|+1)\cdot(\frac{1}{\delta}-1) + \lceil(\log_c(\delta\cdot(1-c)))\rceil \geq \frac{h}{\mu_2}$

$$\begin{aligned}\mathcal{E}_s\left(\left(Y_{T_a^{[a,b]}}\right)_2\right) \geq &\left(\left(\left((k+1)\cdot\left(\frac{1}{\delta}-1\right) + \lceil(\log_c(\delta\cdot(1-c)))\rceil\right)\cdot\mu_2\right)\cdot(1-2g^k) - h\right)\\ &- R\cdot\left(|S| + \frac{2}{1-g}\right)\\ &- R\cdot\frac{2g}{(1-g)^2}\end{aligned}$$

Proof. The first result follows from (8) and Lemmas 26 and 27.

To get the upper bound for $T_b^{[a,b]}$ in general case, we simply add the expected time it takes to reach a BSCC and upper bound the time it takes with the worst possible $Y_{T_C}^{[a,b]}(=a)$. Hence, by Lemmas 20 and 26

$$\mathcal{E}_s\left(T_b^{[a,b]}\right) \leq \mathcal{E}_s(T_C) + \mathcal{E}_q\left(T_b^{[0,b-a]}\right) \leq |S| + \frac{2}{1-g} + \frac{|a|+b+h+R}{\mu}$$

Finally, for the lower bound on $\mathcal{E}_s\left(\left(Y_{T_a^{[a,b]}}\right)_2\right)$, if $\mathcal{E}_s\left(T_a^{[a,b]}\right) = \infty$ we are done as $\mu_2 > 0$ could be a lower bound in this case for achievable mean payoff making $\mathcal{E}_s\left(\left(Y_{T_a^{[a,b]}}\right)_2\right) = \infty$, so assume $\mathcal{E}_s\left(T_a^{[a,b]}\right) < \infty$.

By law of total expectation, partitioning based on whether $T_C < T_a^{[a,b]}$, we have

$$\begin{aligned}\mathcal{E}_s\left(\left(Y_{T_a^{[a,b]}}\right)_2\right) &= \mathcal{E}_s\left(\left(Y_{T_a^{[a,b]}}\right)_2 \,|\, T_a^{[a,b]} < T_C\right) \cdot \mathcal{P}_s\left(T_a^{[a,b]} < T_C\right) \\ &+ \mathcal{E}_s\left(\left(Y_{T_a^{[a,b]}}\right)_2 \,|\, T_a^{[a,b]} \geq T_C\right) \cdot \mathcal{P}_s\left(T_a^{[a,b]} \geq T_C\right)\end{aligned}$$

We will now show lower bounds for each summand separately. For $\mathcal{E}_s\left(\left(Y_{T_a^{[a,b]}}\right)_2 \,|\, T_a^{[a,b]} < T_C\right) \cdot \mathcal{P}_s\left(T_a^{[a,b]} < T_C\right)$, since $r_2(e) \geq -R$ always,

$$\mathcal{E}_s\left(\left(Y_{T_a^{[a,b]}}\right)_2 \,|\, T_a^{[a,b]} < T_C\right) \geq -R \cdot \mathcal{E}_s\left(T_a^{[a,b]} \,|\, T_a^{[a,b]} < T_C\right)$$

$$\begin{aligned}&\mathcal{E}_s\left(T_a^{[a,b]} \,|\, T_a^{[a,b]} < T_C\right) \cdot \mathcal{P}_s\left(T_a^{[a,b]} < T_C\right) \\ &= \sum_{m=k}^{\infty} m \cdot \mathcal{P}_s\left(T_a^{[a,b]} = m \,|\, T_a^{[a,b]} < T_C\right) \cdot \mathcal{P}_s\left(T_a^{[a,b]} < T_C\right) \\ &= \sum_{m=k}^{\infty} m \cdot \mathcal{P}_s\left(T_a^{[a,b]} = m \cap T_a^{[a,b]} < T_C\right) \\ &\leq \sum_{m=k}^{\infty} m \cdot \mathcal{P}_s(T_C > m) \\ &\leq \sum_{m=k}^{\infty} m \cdot 2 \cdot g^m \quad (\textit{Lemma } 26) \\ &\leq \sum_{m=0}^{\infty} m \cdot 2 \cdot g^m \\ &= \frac{2g}{(1-g)^2}\end{aligned}$$

and then using above inequality and (A.3)

$$\mathcal{E}_s\left(\left(Y_{T_a^{[a,b]}}\right)_2 \,|\, T_a^{[a,b]} < T_C\right) \cdot \mathcal{P}_s\left(T_a^{[a,b]} < T_C\right) \geq -R \cdot \frac{2g}{(1-g)^2}$$

For the other summand, we split the sum into two parts; sum of rewards gained until T_C *i.e.*, , until one reaches a BSCC and sum of rewards inside a BSCC. Let $T \overset{\text{def}}{=} T_a^{[a,b]} - T_C$ denote the time spent inside a BSCC and $(Y_T)_2$ denote the sum of rewards inside the BSCC before hitting a. Then by linearity of expectation

$$\mathcal{E}_s\left(\left(Y_{T_a^{[a,b]}}\right)_2 \,|\, T_a^{[a,b]} \geq T_C\right) \cdot \mathcal{P}_s\left(T_a^{[a,b]} \geq T_C\right) = \mathcal{E}_s\left((Y_{T_C})_2 + (Y_T)_2 \,|\, T_a^{[a,b]} \geq T_C\right) \cdot \mathcal{P}_s\left(T_a^{[a,b]} \geq T_C\right)$$

For $(Y_{T_C})_2$, one can follow a similar structure to that of the previous summand. So we first find an upper bound on $\mathcal{E}_s\left(T_C \,|\, T_a^{[a,b]} \geq T_C\right) \cdot \mathcal{P}_s\left(T_a^{[a,b]} \geq T_C\right)$

$$\begin{aligned}\mathcal{E}_s\left(T_C \,|\, T_a^{[a,b]} \geq T_C\right) \cdot \mathcal{P}_s\left(T_a^{[a,b]} \geq T_C\right) &\leq \mathcal{E}_s(T_C) \quad \text{Since } T_C \geq 0 \\ &= \sum_{m=0}^{\infty} \mathcal{P}_s(T_C > m) \\ &= |S| + \frac{2 \cdot g^{|S|}}{1-g} \quad (\textit{Lemma } 26) \\ &\leq |S| + \frac{2}{1-g}\end{aligned}$$

Thus

$$\mathcal{E}_s\Big((Y_{T_C})_2 \,|\, T_a^{[a,b]} \geq T_C\Big) \cdot \mathcal{P}_s\Big(T_a^{[a,b]} \geq T_C\Big) \geq -R \cdot \left(|S| + \frac{2}{1-g}\right).$$

To calculate expectation for $(Y_T)_2$, we condition on the energy level $Y_{T_C}^{[a,b]}$ w.r.t r and the state in which we enter the BSCC. Suppose $Y_{T_C}^{[a,b]} = y$ and $X_{T_C} = q$. But conditioned on $Y_{T_C}^{[a,b]} = y \cap X_{T_C} = q \cap T^{[a,b]_a} \geq T_C$, T is precisely $T_{a-y}^{[a-y,b-y]}$ with $X_0 = q$

$$\begin{aligned}\mathcal{E}_s\Big((Y_T)_2 \,|\, T_a^{[a,b]} \geq T_C \cap X_{T_C} = q \cap \Big(Y_{T_C}^{[a,b]}\Big)_1 = y\Big) &= \mathcal{E}_q\bigg(\Big(Y_{T_{a-y}^{[a-y,b-y]}}\Big)_2 \,|\, T \geq 0\bigg)\\ &= \mathcal{E}_q\bigg(\Big(Y_{T_{a-y}^{[a-y,b-y]}}\Big)_2\bigg)\end{aligned}$$

Also assume the mean payoff w.r.t r_2 in this BSCC is some $\lambda > \mu_2 > 0$. Since T is a stopping time with finite expectation (as per our assumption), we can apply optional stopping theorem to the martingale of r_2 cf. Fact 1.

$$m_{2n}^q = (Y_n)_2 + \nu(X_n) - n\lambda$$

where the index n is actually counted from T_C. This implies

$$\begin{aligned} m_{2T}^q &= m_{20}^q\\ (Y_T)_2 + \nu(X_T) - T\lambda &= \nu(q)\\ (Y_T)_2 &= T\,\lambda + \nu(q) - \nu(X_T)\\ &\geq T\,\mu_2 - h\\ \implies \mathcal{E}_q((Y_T)_2) &\geq \mathcal{E}_q(T)\,\mu_2 - h \qquad = \mathcal{E}_q\Big(T_{a-y}^{[a-y,b-y]}\Big) \cdot \mu_2 - h \end{aligned}$$

Therefore, partitioning over all possible tuples (q, y)

$$\begin{aligned}&\mathcal{E}_s\Big((Y_T)_2 \,|\, T_a^{[a,b]} \geq T_C\Big) \cdot \mathcal{P}_s\Big(T_a^{[a,b]} \geq T_C\Big)\\ &\geq \sum_{(q,y)} \Big(\mathcal{E}_q\Big(T_{a-y}^{[a-y,b-y]}\Big) \cdot \mu_2 - h\Big) \cdot \mathcal{P}_s\Big(T_a^{[a,b]} \geq T_C \cap X_{T_C} = q \cap \Big(Y_{T_C}^{[a,b]}\Big)_1 = y\Big)\end{aligned}$$

When $Y_{T_C}^{[a,b]} \geq -k \cdot R$, then $\mathcal{E}_q\Big(T_{a-y}^{[a-y,b-y]}\Big) \geq (k+1) \cdot \big(\frac{1}{\delta} - 1\big) + \lceil(\log_c(\delta \cdot (1-c)))\rceil$ using Equation (7). Observe that this event subsumes $T_C < k$, so probability of this happening is $\geq 1 - 2g^k$ by Lemma 26. In the other case when $Y_{T_C 1} < -k \cdot R$, a trivial lower bound of 0 for $\mathcal{E}_q\Big(T_{a-y}^{[a-y,b-y]}\Big)$ suffices for our purposes. Since from our assumption, δ is small enough so that $(|S|+1) \cdot \big(\frac{1}{\delta} - 1\big) + \lceil(\log_c(\delta \cdot (1-c)))\rceil \geq \frac{h}{\mu_2}$, putting it all together we have

$$\begin{aligned}&\mathcal{E}_s\Big((Y_T)_2 \,|\, T_a^{[a,b]} \geq T_C\Big) \cdot \mathcal{P}_s\Big(T_a^{[a,b]} \geq T_C\Big)\\ &\geq \left((k+1) \cdot \left(\frac{1}{\delta} - 1\right) + \lceil(\log_c(\delta \cdot (1-c)))\rceil \cdot \mu_2 - h\right) \cdot \big(1 - 2g^k\big) - h \cdot 2g^k\end{aligned}$$

So

$$\begin{aligned}\mathcal{E}_s\Big(\Big(Y_{T_a^{[a,b]}}\Big)_2\Big) \geq{} & -R \cdot \frac{2g}{(1-g)^2}\\ &+ \left(-R \cdot \left(|S| + \frac{2}{1-g}\right)\right)\\ &+ \left(\left(\left((k+1) \cdot \left(\frac{1}{\delta} - 1\right) + \lceil(\log_c(\delta \cdot (1-c)))\rceil\right) \cdot \mu_2\right) \cdot \big(1 - 2g^k\big) - h\right).\end{aligned}$$

◀

B Proof of Lemma 11

▶ **Lemma 11.** *Let $\mu_i > 0$ denote the lower bound on the mean payoff in dimensioni in any BSCC in the Markov chain $\mathcal{A}^{\texttt{Gain}}$ with corresponding computable constants c_i, $g_{\texttt{Gain}}$, $h_{\texttt{Gain}}$ $((5),(9),(1))$, and let μ denote the lower bound on the mean payoff in the 1^{st} dimension in any BSCC of $\mathcal{A}^{\texttt{Bailout}}$ with the corresponding constants $g_{\texttt{Bailout}}, h_{\texttt{Bailout}}$. All the above constants, except c_i, can be chosen as at most exponential in $|\mathcal{M}^*|$ and $1/(1-c_i) \in \mathcal{O}\big(\exp\big(\exp\big(|\mathcal{M}^*|^{\mathcal{O}(1)}\big)\big)\big)$.*

Then there are constants $0 < C_1 < 1$, $C_2 > 0$, $C_3 > 0$, $C_4 > 0$, $C_5 > 0$, all exponential in $|\mathcal{M}^*|$ and dependent only on $\mathcal{M}$, such that for $k \overset{def}{=} \frac{2\cdot|S^*|}{x_{\min,1}^{|S^*|}} \in \mathcal{O}\big(\exp\big(|\mathcal{M}^*|^{\mathcal{O}(1)}\big)\big)$, any $\delta \in (0,1)$ sufficiently small such that $(|S^*|+1)\cdot\big(\frac{1}{\delta}-1\big) + \big\lceil\big(\log_{c_i}(\delta\cdot(1-c_i))\big)\big\rceil \geq \frac{h_{\texttt{Gain}}}{\mu_i}$ for all $2 \leq i \leq d$, one can choose
$Z_g \overset{def}{=} Z_b + R + k{\cdot}R + \max_i\big(R\cdot\lceil\log_{c_i}(\delta\cdot(1-c_i))\rceil - R + 1, h_{\texttt{Gain}}\big) \in \mathcal{O}\Big(e^{|\mathcal{M}^*|^{\mathcal{O}(1)}}\cdot\log(1/\delta)\Big)$ and $b \overset{def}{=} Z_g + 1$ so that

$$\mathcal{E}^{\mathcal{M}^*}_{\sigma^*_{\texttt{alt},\texttt{Z}_\texttt{b},\texttt{Z}_\texttt{g},\texttt{b}}[(Z_g,0)],q}\big((Y_{T^{\texttt{Gain}}})_i\big) \geq C_1\cdot\frac{1}{\delta} - C_2\log_2\left(\frac{1}{\delta}\right) - C_3 \qquad \overset{def}{=} v_i^1$$

$$\mathcal{E}^{\mathcal{M}^*}_{\sigma^*_{\texttt{alt},\texttt{Z}_\texttt{b},\texttt{Z}_\texttt{g},\texttt{b}}[(Z_b,1)],q}\big((Y_{T^{\texttt{Bailout}}})_i\big) \geq -C_4\log_2\left(\frac{1}{\delta}\right) - C_5 \qquad \overset{def}{=} v_i^2$$

In particular, in order to satisfy the condition $v_i^1 + v_i^2 > 0$, it suffices to choose $1/\delta \in \mathcal{O}\big(\max(1/C_1, \max_{2\leq j\leq 5} C_j)^{\mathcal{O}(1)}\big)$. Since the constants C_j are exponential in $|\mathcal{M}^*|$, and by the conditions on the other constants above, the value Z_g, and hence the overall bound $b = Z_g + 1$, can be chosen such that $b \in \mathcal{O}\big(\exp\big(|\mathcal{M}^*|^{\mathcal{O}(1)}\big)\big)$.

Proof. We parametrise $|\mathcal{M}^*|$ along with $\boldsymbol{r}$ on

- Number of states $n \overset{\text{def}}{=} |S^*|$.
- Maximum bit length of probability in P. Let it be w.
- Number of reward dimensions d.
- Maximum reward on an edge in any dimension R.

One can also similarly define size of the Markov chain induced by some finite-memory strategy.

Let $f(n,w,d,R) \overset{\text{def}}{=} n^2(2 + w + d\cdot(1 + \lceil(\log_2(R+1))\rceil))$. Assuming binary representation of rewards, it is easy to see that $|\mathcal{M}| \leq f(n,w,d,R)$. The probabilities are always represented in binary.

As $\sigma^*_{\texttt{Bailout}}$ is MD, $|\mathcal{A}^{\texttt{Bailout}}| \leq f(n,w,d,R)$.
Similarly, as $\sigma^*_{\texttt{Gain}}$ is obtained as a result of a linear program, we have

$$|\mathcal{A}^{\texttt{Gain}}| \leq f(n, LP_{\texttt{Gain}}(n,w,d,R), d, R)$$

where $LP_{\texttt{Gain}}$ (cf. Figure 1) is some fixed polynomial in n, w, d and $\log R$. For succinctness, we define $w_{\texttt{Gain}} \overset{\text{def}}{=} LP_{\texttt{Gain}}(n,w,d,R)$.

To show that all the constants lie in $\mathcal{O}\big(\exp\big(|\mathcal{M}^*|^{\mathcal{O}(1)}\big)\big)$, we simply consider $v_1^i + v_2^i$ and show that each of the constants for $v_1^i + v_2^i$ is in the required size. The result then follows by observing that every constant is positive.

From Lemmas 10 and 28, for

$$Z_g \overset{\text{def}}{=} Z_b + R + k\cdot R + \max\big(R\big\lceil\big(\log_{c_i}(\delta\cdot(1-c_i))\big)\big\rceil + R - 1, h_{\texttt{Gain}}\big)$$

$$\begin{aligned}
&\max \varepsilon && \\
&\sum_{s\in C} \pi_s = 1 && \pi_s : \text{Avg. time spent in } s \\
&\sum_{(s,s')\in E_C} x_{(s,s')} = \pi_s && s \in S_\Box \cap C \\
&x_{(s,s')} = \pi_s \cdot P(s)(s') && s \in S_\bigcirc \cap C \\
&\sum_{(s,s')\in E_C} x_{(s,s')} \cdot \boldsymbol{r}((s,s')) \geq \varepsilon \cdot R && \mathtt{MP}_{[1,d]}(> \mathbf{0}) \\
&\varepsilon \geq 0 &&
\end{aligned}$$

Figure 1 LP for an MEC C for $\mathtt{Gain}$ [7, Figure 3]

it suffices to choose k and δ such that

$$\begin{aligned}
&\left((k+1)\cdot\left(\frac{1}{\delta}-1\right)+\left\lceil\left(\log_{c_i}(\delta\cdot(1-c_i))\right)\right\rceil\right)\cdot\mu_i\cdot\left(1-2g_{\mathtt{Gain}}^{k}\right)- \\
&\qquad h_{\mathtt{Gain}} - R\cdot\left(\left(n+\frac{2}{1-g_{\mathtt{Gain}}}\right)-\frac{2g_{\mathtt{Gain}}}{\left(1-g_{\mathtt{Gain}}\right)^2}\right) > \\
&\qquad R\cdot\left(n+\frac{2}{1-g_{\mathtt{Bailout}}}+\frac{Z+h_{\mathtt{Bailout}}+R}{\mu}\right) \quad \forall\, 2\leq i\leq d \qquad (10) \\
&\qquad k > n \qquad (11) \\
&\left((k+1)\cdot\left(\frac{1}{\delta}-1\right)+\left\lceil\left(\log_{c_i}(\delta\cdot(1-c_i))\right)\right\rceil\right)\cdot\mu_i \geq h_{\mathtt{Gain}} \quad \forall\, 2\leq i\leq d \qquad (12)
\end{aligned}$$

The last set of equations become redundant due to the first one. The left-hand side of (10) is the over precision of constant v_1^i and similarly the right-hand side is that of $-v_i^2$. It is simple to notice that once k is fixed, then v_1^i varies with δ as $C_1\cdot\frac{1}{\delta}-C_2\log\left(\frac{1}{\delta}\right)-C_3$ and v_2^i as $-C_4\log\left(\frac{1}{\delta}\right)-C_5$ for some appropriate constants C_i. To further simplify, W.l.o.g, assume δ is sufficiently small such that

$$R\left\lceil\left(\log_{c_i}(\delta\cdot(1-c_i))\right)\right\rceil + R \geq h_{\mathtt{Gain}}.$$

By definition of Z_g and our assumption on δ, we get that

$$Z_g = Z_b + R + k\cdot R + R\left\lceil\left(\log_{c_i}(\delta\cdot(1-c_i))\right)\right\rceil + R - 1.$$

Then rearranging constants to one side and terms depending on k and δ on other side we get

$$\begin{aligned}\left((k+1)\cdot\left(\frac{1}{\delta}-1\right)+\lceil(\log_{c_i}(\delta\cdot(1-c_i)))\rceil\right)\cdot\mu_i\cdot(1-2g^k_{\texttt{Gain}})&\\ -(k+\lceil(\log_{c_i}(\delta\cdot(1-c_i)))\rceil)\cdot\frac{R^2}{\mu}&>\\ h_{\texttt{Gain}}+R\cdot\left(\left(n+\frac{2}{1-g_{\texttt{Gain}}}\right)+\frac{2\cdot g_{\texttt{Gain}}}{(1-g_{\texttt{Gain}})^2}\right)&\\ +R\cdot\left(n+\frac{2}{1-g_{\texttt{Bailout}}}+\frac{Z_b+3\cdot R-1+h_{\texttt{Bailout}}}{\mu}\right)&\end{aligned}$$

We will upper bound the RHS and lower bound LHS by simpler formulas to get sufficient bounds on k and δ. Let $x_{\min,1}$ denote the minimum probability in $\mathcal{A}^{\texttt{Gain}}$, and $x_{\min,2}$ denote the minimum probability in $\mathcal{A}^{\texttt{Bailout}}$. Then by definition of $w_{\texttt{Gain}}$ and w,

$$x_{\min,1}\geq\frac{1}{2^{w_{\texttt{Gain}}}}\qquad x_{\min,2}\geq\frac{1}{2^{w}}$$

from (1)

$$\begin{aligned}h_{\texttt{Gain}}&=\frac{2\cdot n\cdot R}{x^n_{\min,1}}\\ &\leq 2^{1+\lceil\log_2 n+\log_2 R\rceil+n\cdot w_{\texttt{Gain}}}\\ &\leq 2^{f(n,w_{\texttt{Gain}},d,R)}\end{aligned}$$

Similarly, one gets $h_{\texttt{Bailout}}\leq 2^{f(n,w,d,R)}$.

From (9)

$$\begin{aligned}1-g_{\texttt{Gain}}&=1-\exp\left(\frac{-x^n_{\min,1}}{n}\right)\\ &\geq\frac{x^n_{\min,1}}{2n}\left(\text{Since } 1-e^{-x}\geq\frac{e-1}{e}\cdot x\geq\frac{x}{2}\text{ for } x\in[0,1]\right) && (13)\\ \implies R\cdot\frac{2}{1-g_{\texttt{Gain}}}&\leq\frac{4\cdot n\cdot R}{x^n_{\min,1}} && (14)\\ &=2h_{\texttt{Gain}} && (15)\\ &\leq 2\,2^{f(n,w_{\texttt{Gain}},d,R)} && (16)\end{aligned}$$

Similarly, $R\cdot\frac{2}{1-g_{\texttt{Bailout}}}\leq 2\,2^{f(n,w,d,R)}$

$$\begin{aligned}R\cdot\frac{2g_{\texttt{Gain}}}{(1-g_{\texttt{Gain}})^2}&\leq R\cdot\frac{2}{(1-g_{\texttt{Gain}})^2}\\ &\leq\frac{2\cdot n^2\cdot R}{x^{2n}_{\min,1}}\\ &\leq 2^{1+\lceil 2\log n+\log R\rceil+2n\cdot w_{\texttt{Gain}}}\\ &\leq 2^{2f(n,w_{\texttt{Gain}},d,R)}\end{aligned}$$

To get a lower bound on μ, let $\mathcal{B}$ be any BSCC of $\mathcal{A}^{\texttt{Bailout}}$ and $P_{\mathcal{B}}$ be the one-step transition probability matrix in $\mathcal{A}^{\texttt{Bailout}}$ restricted to $\mathcal{B}$. Clearly the number of states in $\mathcal{B}$ is $\leq n$. The steady state probabilities $\pi_{\mathcal{B}}$ are solution to the linear system Figure 2.

$$\begin{aligned}
(I - P_{\mathcal{B}})^T \cdot \pi_{\mathcal{B}} &= \mathbf{0} \\
\sum_{s \in \mathcal{B}} \pi_{\mathcal{B}}(s) &= 1 \\
\pi_{\mathcal{B}} &\geq \mathbf{0}
\end{aligned}$$

Figure 2 LP$_1$: Linear program for steady state probabilities in a BSCC

We apply [23, Theorem 15]. First, lets multiply each row by lcm of denominators to get integer entries. The size of each entry is now bounded by $n \cdot w$. Therefore, size of the entire matrix is $\leq n^3 \cdot w$. $size(b)$ here $\leq 2n+2$. $\implies$ the denominator of each component of $\pi_{\mathcal{B}}$ is $\leq 2^{\left(n^3 w + 2n + 2\right)} \leq 2^{f^2(n,w,d,R)}$.

The mean payoff in this BSCC is then given by

$$\mu_{\mathcal{B}} \overset{\text{def}}{=} \sum_{s} \sum_{\{s' \mid (s,s') \in E_{\mathcal{B}}\}} \pi_{\mathcal{B}}(s) \cdot P(s)(s') \cdot r_1((s,s'))$$

The least common denominator for all such $P(s)(s')$ will be $\leq 2^{n \cdot w} \leq 2^{f(n,w,d,R)}$ which means the overall denominator for $\mu_{\mathcal{B}} \leq 2^{f^2(n,w,d,R)+f(n,w,d,R)} \leq 2^{2f^2(n,w,d,R)}$.

Therefore, $\mu_{\mathcal{B}} \geq 2^{-2f^2(n,w,d,R)}$. Since μ is just minimum over all such $\mu_{\mathcal{B}}$,

$$\mu \geq 2^{-2f^2(n,w,d,R)}$$

Finally, $Z_b = \max_s i_s^{\mathtt{Bailout}} \leq 3 \cdot n \cdot R$. Combining everything and from the fact that $w_{\mathtt{Gain}} \geq w$, we get that RHS

$$\begin{aligned}
&\leq 2^{f(n,w_{\mathtt{Gain}},d,R)} + 2 \cdot 2^{f(n,w_{\mathtt{Gain}},d,R)} + 2^{2f(n,w_{\mathtt{Gain}},d,R)} + 2 \cdot n \cdot R \\
&+ 2 \cdot 2^{f(n,w,d,R)} + R^2 \cdot 2^{2f^2(n,w,d,R)} \left(3 \cdot n \cdot R + 3 \cdot R + 2^{3f^2(n,w,d,R)}\right) \\
&\leq 5 \cdot 2^{f(n,w_{\mathtt{Gain}},d,R)} + 2 \cdot n \cdot R + 2^{2f(n,w_{\mathtt{Gain}},d,R)} \\
&+ R^2 \cdot 2^{2f^2(n,w,d,R)} \left(3 \cdot (n+1) \cdot R + 2^{3f^2(n,w,d,R)}\right) \\
&\leq 7 \cdot 2^{2f(n,w_{\mathtt{Gain}},d,R)} + 2^{2f(n,w,d,R)} \cdot 2^{2f^2(n,w,d,R)} \left(2^{2f(n,w,d,R)} + 2^{3f^2(n,w,d,R)}\right) \\
&\leq 2^{2f(n,w_{\mathtt{Gain}},d,R)+3} + 2^{4f^2(n,w,d,R)} \cdot 2^{5f^2(n,w,d,R)} \\
&\leq 2 \cdot 2^{9f^2(n,w_{\mathtt{Gain}},d,R)}
\end{aligned}$$

To lower bound LHS, first choose k to be sufficiently large such that $g_{\mathtt{Gain}}^k \leq 1/4$. Let $k = \lceil \frac{2n}{x_{\min,1}^n} \rceil \geq n+1$. Then

$$\begin{aligned}
\left((k+1) \cdot \left(\frac{1}{\delta} - 1\right) + \lceil (\log_{c_i}(\delta \cdot (1 - c_i))) \rceil \right) \cdot \mu_i \cdot (1 - 2g_{\mathtt{Gain}}^k) &\geq (k+1) \cdot \left(\frac{1}{\delta} - 1\right) \cdot \frac{\mu_i}{2} \\
&\geq \frac{k\mu_i}{2\delta} \quad \text{Assume } \delta < \frac{1}{k+1} \\
&\geq 2^{-2f^2(n,w_{\mathtt{Gain}},d,R)} \cdot \frac{1}{\delta} \\
&\text{since } k \geq 2 \text{ and } \mu_i \geq 2^{-2f^2(n,w_{\mathtt{Gain}},d,R)}
\end{aligned}$$

$$
\begin{aligned}
\frac{k\,R^2}{\mu} &\leq k\,R^2\ 2^{f(n,w,d,R)}\\
&\leq 2^{2\,f(n,w_{\mathtt{Gain}},d,R)}\cdot 2^{f(n,w,d,R)}\\
\implies -\frac{k\,R^2}{\mu} &\geq 2^{3\,f(n,w_{\mathtt{Gain}},d,R)}
\end{aligned}
$$

From (5) and (4)

$$
\begin{aligned}
\log_{c_i}(\delta\cdot(1-c_i))\cdot\frac{R^2}{\mu} &= \frac{\log 1/\delta+\log(1/(1-c_i))}{\log 1/c_i}\cdot\frac{R^2}{\mu}\\
&\leq (\log 1/\delta+\log(1/(1-c_i)))\cdot\frac{2\eta_i^2\cdot R^2}{\mu_i^2\cdot\mu}
\end{aligned}
$$

Using $\eta_i \leq 3\cdot h_{\mathtt{Gain}}$, and $1-c_i \geq \frac{\mu_i^2}{2\eta_i^2}$, we get

$$
\begin{aligned}
&\leq (\log_2(1/\delta)+1+2\cdot\log_2\eta_i+2\cdot\log_2 1/\mu_i)\cdot(3h_{\mathtt{Gain}}R)^2\cdot 2^{6f^2(n,w_{\mathtt{Gain}},d,R)}\\
&\leq \big(\log_2(1/\delta)+1+2\cdot\log_2 3+2f(n,w_{\mathtt{Gain}},d,R)+4f^2(n,w_{\mathtt{Gain}},d,R)\big)\cdot 9\cdot 2^{10f^2(m,w_{\mathtt{Gain}},d,R)}\\
&\leq \big(\log_2(1/\delta)+5+6\cdot f^2(n,w_{\mathtt{Gain}},d,R)\big)\cdot 9\cdot 2^{10f^2(n,w_{\mathtt{Gain}},d,R)}\\
&\leq 9\cdot 2^{10f^2(n,w_{\mathtt{Gain}},d,R)}\log_2\left(\frac{1}{\delta}\right)+99\cdot 2^{11f^2(n,w_{\mathtt{Gain}},d,R)}
\end{aligned}
$$

Combining everything, we have LHS

$$
\geq 2^{-2f^2(n,w_{\mathtt{Gain}},d,R)}\cdot\frac{1}{\delta}-9\cdot 2^{10f^2(n,w_{\mathtt{Gain}},d,R)}\log_2\left(\frac{1}{\delta}\right)-992^{11f^2(n,w_{\mathtt{Gain}},d,R)}-2^{3f(n,w_{\mathtt{Gain}},d,R)} \quad (17)
$$

Comparing it with the constants from Lemma 11, one can see that $C_1 = 2^{-2f^2(n,w_{\mathtt{Gain}},d,R)}$, $C_2+C_3 = 9\cdot 2^{10f^2(n,w_{\mathtt{Gain}},d,R)}$ and $C_4+C_5 = 102\,2^{11f^2(n,w_{\mathtt{Gain}},d,R)}$ all of which are in the required complexity.

Finally, it suffices to choose a δ such that

$$
2^{-2f^2(n,w_{\mathtt{Gain}},d,R)}\cdot\frac{1}{\delta}-9\cdot 2^{10f^2(n,w_{\mathtt{Gain}},d,R)}\log_2\left(\frac{1}{\delta}\right) > 102\,2^{11f^2(n,w_{\mathtt{Gain}},d,R)}
$$

$\delta = 2^{-20f^2(n,w_{\mathtt{Gain}},d,R)}$ should satisfy the required inequality. Therefore, with $k = \lceil\frac{2n}{x_{\min,1}^n}\rceil$, and $\delta = 2^{-20f^2(n,w_{\mathtt{Gain}},d,R)}$ the overall bound b will be exponential in $|\mathcal{M}|$. ◀

C Proofs of Section 5

Before proceeding with the proof of Lemma 13, we state some useful definitions and prove some intermediate lemmas which makes it easier to understand the idea. We start by defining the notion of a winning end component (WEC).

▶ **Definition 29.** *Let $\mathcal{M}(B) = (S_B, S_{\Box B}, S_{\bigcirc B}, E_B, \boldsymbol{r}_B)$ be an end component of $\mathcal{M}$. We say that $\mathcal{M}(B)$ is a WEC (winning end component) iff there is some strategy $\sigma \in \Sigma_f^{\mathcal{M}(B)}$ such that*

- *$\mathcal{M}(B)^\sigma$ is irreducible and the end component defined by it is exactly $\mathcal{M}(B)$.*
- *for every state $q \in S_B$, there is some minimal energy level j_q such that $\mathcal{P}_{\sigma,q}^{\mathcal{M}(B)}(\mathtt{0}(j_q)) = 1$.*

We simply say B is a WEC instead of $\mathcal{M}(B)$ is a WEC for succinctness.

Furthermore, denote by $\mathbb{C}(B) \overset{\text{def}}{=} \{\mathsf{C} \mid \mathsf{C} \text{ is a simple cycle in } \mathcal{M}(B)\}$, the set of all simple cycles in $\mathcal{M}(B)$ and given a simple cycle $\mathsf{C} = s_0 \xrightarrow{c_0} s_1 \xrightarrow{c_1} \ldots s_j = s_0$ be a cycle of length j, where c_i denotes the rewards in the energy (1^{st}) dimension, define the effect of C to be $\mathtt{eff}(\mathsf{C}) \overset{\text{def}}{=} \sum_{k=0}^{j-1} c_k$.

A WEC B is called a WEC of Type-I if there is some $\mathsf{C} \in \mathbb{C}(B)$ such that $\mathtt{eff}(\mathsf{C}) > 0$. Otherwise, it is called a WEC of Type-II.

▶ **Lemma 30.** *If B is a WEC of Type-I, then one can choose σ such that it satisfies all the conditions of Definition 29 along with*

$$\mathcal{P}_{\sigma,q}^{\mathcal{M}(B)}(\mathtt{MP}_1(>0)) = 1$$

for every state $q \in S_B$.

Proof. Assume that $\sigma = (\mathsf{M}, \mathsf{m}_0, \mathsf{upd}, \mathsf{nxt})$ which satisfies the requirements of Definition 29 gives a mean payoff of 0 in the energy dimension. Let $\mathfrak{C} = (s_0, \mathsf{m}_0) \xrightarrow{c_0} (s_1, \mathsf{m}_1) \xrightarrow{c_1} \ldots (s_k, \mathsf{m}_k) = (s_0, \mathsf{m}_k)$ be a simple cycle of length k in $\mathcal{M}(B)^{\sigma}$.

▷ Claim 31. For any cycle $\mathfrak{C}$ in $\mathcal{M}^{\sigma}$, $\mathtt{eff}(\mathfrak{C}) = 0$.

Proof. We have $\mathcal{P}_{\sigma,s}^{\mathcal{M}(B)}(\mathtt{O}(j_s)) = 1$, and $\mathcal{P}_{\sigma,s}^{\mathcal{M}}(\mathtt{MP}_1(>0)) = 0$. The former implies that $\mathtt{MP}_1(\geq 0)$ surely. In fact, it can be never be the case that $\mathtt{eff}(\mathfrak{C}) < 0$ as otherwise $\mathtt{EN}_1$ and hence $\mathtt{O}(j_s)$ is not satisfied almost surely. If $\mathtt{eff}(\mathfrak{C}) > 0$ for some $\mathfrak{C}$, this then implies a positive mean payoff since $\mathcal{M}^{\sigma}$ is an irreducible, finite Markov chain, a contradiction. Hence, $\mathtt{eff}(\mathfrak{C}) = 0$. ◀

We construct a strategy σ' which follows Definition 29 such that $\mathcal{P}_{\sigma',s}^{\mathcal{M}(B)}(\mathtt{O}(j_s) \cap \mathtt{MP}_1(>0)) = 1$. Since B is a WEC of Type-I, there is some cycle C with positive effect. And since every edge in E_B occurs somewhere in $\mathcal{M}^{\sigma}$ by definition of σ, there is some cycle $\mathfrak{C}$ such that $(s_{k_1}, \mathsf{m}_{k_1}) \xrightarrow{c_{k_1}, x_{k_1}} (s_{k_1+1}, \mathsf{m}_{k_1+1}) \xrightarrow{c_{k_1+1}, x_{k_1+1}} \ldots (s_{k_1+|\mathsf{C}|}, \mathsf{m}_{k_1+|\mathsf{C}|}) \in s_{k_1} \times \mathsf{M} \setminus \mathsf{m}_{k_1}$ is a part of $\mathfrak{C}$ corresponding to C. The strategy σ' behaves exactly like σ except on this specific path at $(s_{k_1+|\mathsf{C}|-1}, \mathsf{m}_{k_1+|\mathsf{C}|-1})$, where instead of continuing in the path of $\mathfrak{C}$ and going to $(s_{k_1}, \mathsf{m}_{k_1+|\mathsf{C}|})$ with probability $x_{k_1+|\mathsf{C}|-1}$, this probability is split into two parts. For some sufficiently small $\varepsilon > 0$

- with probability $(1-\varepsilon) \cdot x_{k_1+|\mathsf{C}|-1}$ it goes to $(s_{k_1}, \mathsf{m}_{k_1+|\mathsf{C}|})$
- with probability $\varepsilon \cdot x_{k_1+|\mathsf{C}|-1}$ it returns to $(s_{k_1}, \mathsf{m}_{k_1})$

Observe that it doesn't matter if $s_{k_1+|\mathsf{C}|-1}$ was a random or a controlled state as the final destination for both edges is the same with only the memory mode being different, so σ' is a valid strategy which updates its memory stochastically. σ' should now not only remember the memory state but also the path in which it arrived at that particular combination, but this can be done with finite memory.

It is easy to see that σ' also induces an irreducible Markov chain with every edge in E_B occurring at least once, and that the energy objective $\mathtt{EN}(j_s)$ is satisfied as the only additional loop which occurs in $\mathcal{M}^{\sigma'}$ has a positive effect on the energy level. Furthermore, for sufficiently small ε, it doesn't change the mean payoff in other dimensions by much thereby still ensuring that σ' satisfies $\mathtt{MP}_{[2,d]}(> \mathbf{0})$. Finally, the addition of this new loop now causes the mean payoff in 1^{st} dimension to be strictly > 0. ◀

Lemma 30 shows that it is possible to win both $\mathtt{Gain}$ and $\mathtt{Bailout}$ almost surely in $\mathcal{M}(B)$ from every state in $q \in B$ whenever B is a Type-I WEC. I.e., $q \in \mathtt{AS}^{\mathcal{M}(B)}\big(\mathtt{MP}_{[1,d]}(> \mathbf{0})\big)$.

Moreover, the minimal safe energy for $\texttt{Bailout}$ in $\mathcal{M}(Q)$ from q is exactly j_q, that is $q \in \texttt{AS}^{\mathcal{M}(B)}(\texttt{EN}_1(j_q) \cap \texttt{MP}_1(>0))$. Thus, $\mathcal{M}(B)$ satisfies the conclusion of Lemma 7, i.e., it behaves like $\mathcal{M}^*$.

Therefore, the strategy $\sigma^*_{\texttt{alt},\texttt{Z}_\texttt{b},\texttt{Z}_\texttt{g},\texttt{b}}$, defined in Section 4, is almost surely winning $\texttt{O}(j_q)$ in $\mathcal{M}(B)$. We can now carry over the analysis on the memory bound b for $\mathcal{M}^*$ from Lemmas 10 and 11 to $\mathcal{M}(B)$. The only difference is that the size is now measured in $|\mathcal{M}(B)| \leq |\mathcal{M}|$. So we obtain the following lemma.

▶ **Lemma 32.** *If B is a WEC of Type-I, there exists a bound $b_B = \mathcal{O}\big(\exp(|\mathcal{M}(Q)|^{\mathcal{O}(1)})\big)$ such that for all $q \in B$*

$$\mathcal{P}^{\mathcal{M}(B)}_{\sigma^*_{\texttt{alt},\texttt{Z}_\texttt{b},\texttt{Z}_\texttt{g},\texttt{b}_\texttt{B}},q}(\texttt{O}(j_q)) = 1.$$

By Remark 12, it is also true that $\mathcal{P}^{\mathcal{M}(B)}_{\sigma^*_{\texttt{alt},\texttt{Z}_\texttt{b},\texttt{Z}_\texttt{g},\texttt{b}_\texttt{B}},q}(\texttt{O}(j_q) \cap \texttt{Infix}(b_B)) = 1$.

Note that, if there is no positive effect cycle in B, there cannot be any cycle with negative effect as well since every cycle is taken infinitely often in a WEC. So, in contrast to Type-I, if B is such that $\texttt{eff}(\mathsf{C}) = 0$ for every simple cycle C, then B is called a WEC of Type-II. But this implies that the maximum fluctuation in energy level is at most $|S_B| \cdot R \leq |S| \cdot R$. Therefore, we get the following.

▶ **Lemma 33.** *For every WEC B of Type-II, there is a finite-memory strategy σ_B with a constant $b_B \in \mathcal{O}(|S| \cdot R)$ such that*

$$\mathcal{P}^{\mathcal{M}(B)}_{\sigma_B,s}(\texttt{O}(j_s) \cap \texttt{Infix}(b_B)) = 1.$$

We are now ready to prove Lemma 13.

▶ **Lemma 13.** *Let B be a BSCC of $\mathcal{M}^\sigma$ and let $\mathcal{M}(B)$ be the corresponding end component in $\mathcal{M}$ with states S_B and transitions E_B. Then there is a strategy σ_B, a bound $b_B \in \mathcal{O}\big(\exp(|\mathcal{M}(B)|^{\mathcal{O}(1)})\big)$ such that for any state $q \in S_B$, there is a minimal safe energy level $j_q \overset{\mathit{def}}{=} i_q^{\mathcal{M}(B)} \leq 3 \cdot |S_B| \cdot R$ such that $\mathcal{P}^{\mathcal{M}(B)}_{\sigma_B,q}(\texttt{O}(j_q) \cap \texttt{Infix}(b_B)) = 1$.*

Proof. We provide the bounds based on the type of the end component $\mathcal{M}(B)$. First observe that B is a WEC as σ acts as a witness by satisfying the requirements of Definition 29. If B is a WEC of Type-II, then by Lemma 33 the minimal energy j_q required to win from any state q in S_B is $\leq |S_B| \cdot R \leq |S| \cdot R$ and the constant b_B is bounded by $\mathcal{O}(|S| \cdot R)$. Choose σ_B and b_B be the strategy and the constant from Lemma 33 in this case. Otherwise, B is a WEC of Type-I. Therefore, j_q in this case would be the same as the minimal energy to satisfy $\texttt{Bailout}$ which by Lemma 5 is $\leq 3|S_B|R \leq 3|S|R$. By choosing σ_B as $\sigma^*_{\texttt{alt},\texttt{Z}_\texttt{b},\texttt{Z}_\texttt{g},\texttt{b}_\texttt{B}}$, with b_B from Lemma 32, we are done. ◀

▶ **Lemma 14.** *Let T denote the union of all S_B of every BSCC B of $\mathcal{M}^\sigma$, as in Lemma 13. Then one can almost surely reach any state in T with the corresponding minimal safe energy level with energy fluctuations of at most $5 \cdot |S| \cdot R$.*

Proof. It is clear that σ is also a witness for $\texttt{EN}_1(i_s) \cap \mathsf{F}T$. But by [13, Lemma 2], this can be achieved with at most $2|S|R$ fluctuation in energy. However, since we also need to ensure that the necessary minimal energy level, one can then simply encode the energy level into the state space of $\mathcal{M}$ and enlarge $\mathcal{M}$ up to $(3+2)|S|R$. So the states of this new MDP $\mathcal{M}'$ will now be $S \times [0, 5|S|R]$. Let $T' = \bigcup_{q \in S_B} q \times [i_q^B, 5|S|R]$. Then, it is not hard to see that when starting from (s, i_s) in $\mathcal{M}'$, one almost surely satisfies $\mathsf{F}T'$. (Move according to

σ until you hit the maximum energy level of $5|S|R$, at which point switch to one of the winning strategies for $\mathtt{EN}_1(\cdot) \cap \mathsf{F}\,T$ which uses only $\mathcal{O}(|S| \cdot R)$ memory modes.) Therefore, one can reach a state q in a BSCC with its safe energy level with a fluctuation of at most $5 \cdot |S| \cdot R$. ◀