Optimizing the Expected Mean Payoff in Energy Markov Decision Processes

Tomáš Brázdil[1], Antonín Kučera[1], and Petr Novotný[2]*

[1] Faculty of Informatics MU, Botanická 68a, 602 00 Brno, Czech Republic,
{brazdil,kucera}@fi.muni.cz
[2] IST Austria, Klosterneuburg, Austria, petr.novotny@ist.ac.at

Abstract. Energy Markov Decision Processes (EMDPs) are finite-state Markov decision processes where each transition is assigned an integer counter update and a rational payoff. An EMDP configuration is a pair $s(n)$, where s is a control state and n is the current counter value. The configurations are changed by performing transitions in the standard way. We consider the problem of computing a safe strategy (i.e., a strategy that keeps the counter non-negative) which maximizes the expected mean payoff.

1 Introduction

Resource-aware systems are systems that consume/produce a discrete resource, such as (units of) time, energy, or money, along their runs. This resource is *critical*, i.e., if it is fully exhausted along a run, a severe runtime error appears and such a situation should be avoided to the largest possible extent. Technically, resource-aware systems are modeled as finite-state programs operating over an integer counter representing the resource. A *configuration* is a pair $s(n)$ where s is the current control state and n is the number of currently available resource units. Each transition is assigned an integer *update* modeling the consumption/production of the resource caused by performing the transition.

Our Contribution. In this paper, we concentrate on the *long-run average optimization problem* for resource-aware systems with both controllable and stochastic states. That is, we assume that the finite control of our resource-aware system is a finite-state Markov decision process (MDP), and each transition is assigned (in addition to the integer counter update) a rational *payoff*[3]. The resulting model is called *energy Markov decision process (EMDP)*. Intuitively, given an EMDP and its initial configuration, the task is to compute a *safe* strategy maximizing the *expected mean payoff*. Here, a strategy is safe if it ensures that the counter stays non-negative along all runs. The *value* of a given configuration $s(n)$, denoted by $Val(s(n))$, is the supremum of all expected mean

* The research has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement no [291734].

[3] The payoff may correspond to some independent performance measure, or it can reflect the use of the critical resource represented by the counter.

payoffs achievable by a safe strategy, and a strategy is *optimal* for $s(n)$ if it is safe and achieves the value. Observe that $Val(s(n)) \geq Val(s(m))$ whenever $n \geq m$, and hence we can also define the *limit value* of s, denoted by $Val(s)$, as $\lim_{n\to\infty} Val(s(n))$.

Since optimal safe strategies may not exists in general, the first natural question is the following:

[Q1]. *Can we determine a "reasonable" condition under which an optimal strategy exists?*

By "reasonable" we mean that the condition should be decidable (with low complexity) and tight (i.e., we should provide counterexamples witnessing that optimal strategies do not necessarily exist if the condition is violated). Further, there are two basic algorithmic questions.

[Q2]. *Can we compute $Val(s(n))$ for a given configuration $s(n)$? If not, can we at least approximate the value up to a given absolute error $\varepsilon > 0$? Can we compute/approximate $Val(s)$ for a given state s? What is the complexity of these problems?*

To show that computing an ε-approximation of $Val(s(n))$ is computationally hard, we consider the following *gap threshold problem*: given a configuration $t(k)$ of a given EMDP and numbers x, ε, where $\varepsilon > 0$, such that either $Val(t((k)) \geq x$ or $Val(t(k)) \leq x-\varepsilon$, decide which of these two alternatives holds[4]. Note that if the gap threshold problem is X-hard for some complexity class X, then $Val(s(n))$ cannot be ε-approximated in polynomial time unless $\mathsf{X} = \mathsf{P}$.

[Q3]. *Can we compute (a finite description of) an optimal strategy for a given configuration (if it exists)? For a given $\varepsilon > 0$, can we compute an ε-optimal strategy? How much memory is required by these strategies? What is the complexity of the strategy synthesis problems?*

Before formulating our answers to the above questions, we need to briefly discuss the relationship between EMDPs and *energy games* [16,15,4].

The problems of **[Q2]** and **[Q3]** subsume the question whether a given configuration of a given EMDP is safe. This problem can be solved by algorithms for 2-player non-stochastic energy games [14], where we treat the stochastic vertices as if they were controlled by an adversarial player. The correctness of this approach stems from the fact that keeping the energy level non-negative is an objective whose violation is witnessed by a finite prefix of a run. Let EG (Energy Games) be the problem of deciding whether a given configuration in a given energy game is safe. A P^{EG} *algorithm* is a deterministic polynomial-time algorithm which inputs an EMDP $\mathcal{E}$ (and possibly some initial configuration $s(n)$ of $\mathcal{E}$) and uses an oracle which freely decides the safety problem for the configurations of $\mathcal{E}$. We assume that the counter updates and rewards used in $\mathcal{E}$, and the n in $s(n)$, are encoded as (fractions of) binary numbers. The size of $\mathcal{E}$ and $s(n)$ is denoted by $|\mathcal{E}|$ and $|s(n)|$, respectively. It is known that EG is solvable in pseudo-polynomial time, belongs to $\mathsf{NP} \cap \mathsf{coNP}$, and it is at least as hard as the parity game problem. From this we immediately obtain that every decision problem solvable by a P^{EG} algorithm belongs to $\mathsf{NP} \cap \mathsf{coNP}$, and every P^{EG} algorithm runs in pseudo-polynomial time, i.e., in time polynomial in $|\mathcal{E}|$, $|s(n)|$, and $M_{\mathcal{E}}$, where $M_{\mathcal{E}}$ is the maximal absolute value

[4] Formally, the decision algorithm answers "yes" iff the first (or the second) possibility holds.

of a counter update in $\mathcal{E}$. We say that a decision problem X is EG-*hard* if there is a polynomial-time reduction from EG to X.

Our results (answers to **[Q1]**–**[Q3]**) can be formulated as follows:

[A1]. We show that an optimal strategy is guaranteed to exist in a configuration $s(n)$ if the underlying EMDP is *strongly connected and pumpable*. An EMDP is strongly connected if its underlying graph is strongly connected, and pumpable if for every safe configuration $t(m)$ there exists a safe strategy σ such that the counter value is unbounded in almost all runs initiated in $t(m)$.

The problem whether a given EMDP is strongly connected and pumpable is in P^{EG} and EG-hard. Further, an optimal strategy in $s(n)$ does not necessarily exist if just one of these two conditions is violated. We use SP-EMDP to denote the subclass of strongly connected and pumpable EMDPs.

[A2, A3]. If a given EMDP belongs to the SP-EMDP subclass, the following holds:

- The value of every safe configuration is the same and computable by a P^{EG} algorithm (consequently, the limit value of all states is also the same and computable by a P^{EG} algorithm). The gap threshold problem is EG-hard.
- There exists a strategy σ which is optimal in every configuration. In general, σ may require infinite memory. A finite description of σ is computable by a P^{EG} algorithm. The same holds for ε-optimal strategies where $\varepsilon > 0$, except that ε-optimal strategies require only finite memory.

Note that since the gap threshold problem is EG-hard, approximating the value is not much easier than computing the value precisely for SP-EMDPs.

For general EMDPs, optimal strategies are not guaranteed to exist. Still, for every EMDP $\mathcal{E}$ we have the following:

- The value of every configuration $s(n)$ can be approximated up to an arbitrarily small given $\varepsilon > 0$ in time polynomial in $|\mathcal{E}|$, $|s(n)|$, $M_{\mathcal{E}}$, and $1/\varepsilon$. The limit value of each control state is computable in time polynomial in $|\mathcal{E}|$ and $M_{\mathcal{E}}$.
- For a given $\varepsilon > 0$, there exists a strategy σ which is ε-optimal in every configuration. In general, σ may require infinite memory. A finite description of σ is computable in time polynomial in $|\mathcal{E}|$, $M_{\mathcal{E}}$, and $1/\varepsilon$.
- The gap threshold problem is PSPACE-hard.

The above results are non-trivial and based on detailed structural analysis of EMDPs. As a byproduct, we yield a good intuitive understanding on what can actually happen when we wish to construct a (sub)optimal strategy in a given EMDP configuration. The main steps are sketched below (we also try to explain where and how we employ the existing ideas, and where we needed to invent original techniques). The details and examples illustrating the discussed phenomena are given later in Section 3.

The core of the problem is the analysis of maximal end components of a given EMDP, so let us suppose that our EMDP is strongly connected (but not necessarily pumpable). First, we check whether there exists *some* strategy such that the average change of the counter per transition is positive (this can be done by linear programming) and distinguish two possibilities:

If there is such a strategy, then we try to optimize the mean payoff under the constraint that the average change of the counter is non-negative. This can be formulated by a linear program whose solution allows to construct finitely many randomized memoryless strategies and an appropriate "mixing ratio" for these strategies that produces an optimal mean payoff. This part is inspired by the technique used in [6] for the analysis of MDPs with multiple mean-payoff objectives. However, here we cannot implement the optimal mixing ratio "immediately" because we also need to ensure that the resulting strategy is safe. We can solve this problem using two different methods, depending on whether the EMDP is pumpable or not. If it is not pumpable, then, since we aim at constructing an ε-optimal strategy, we can always slightly modify the mix, adding the aforementioned strategy which increases the counter in a right proportion. If the counter becomes too low, we permanently switch to some safe strategy (which may produce a low mean payoff). Since the counter has a tendency to increase, we can setup everything so that the probability of visiting low counter values is very small if we start with a sufficiently large initial counter value. Hence, for configurations with a sufficiently large counter value, we play ε-optimally. For the configurations with "low" counter value, we compute a suboptimal strategy by "cutting" the counter when it reaches a large value (where we already know how to play) and applying the algorithm for finite-state MDPs.

More interesting is the case when the EMDP *is* pumpable. Here, instead of switching to *some* safe strategy, we switch to a *pumping* strategy, i.e. a safe strategy that is capable of increasing the counter above any threshold with probability 1. Once the pumping strategy increases the counter to some sufficiently high value, we can switch back to playing the aforementioned "mixture." To obtain an optimal strategy in this way, we need to extremely carefully set up the events which trigger "(de-)activation" of the pumping strategy, so as to ensure that it keeps the counter sufficiently high and at the same time assure that it does not negatively affect the mean payoff. We innovatively use the martingale techniques designed in [8] to accomplish this delicate task.

If there is no such strategy, we need to analyze our EMDP differently. We prove that *every* safe strategy then satisfies the following: almost all runs end by an infinite suffix where all visited configurations with the same control state have the same counter value. This implies that only finitely many configurations are visited in the suffix, and we can analyze the associated mean payoff by methods for finite-state MDPs.

If we additionally assume that our strongly connected EMDP is pumpable, than there inevitably exists a strategy which increases the counter on average (which rules out the second possibility mentioned above) and the "switching" strategy can be constructed differently so that it achieves the optimal mean payoff specified by the linear program.

Let us note that some of the presented ideas can be easily extended even to multi-energy MDPs. Since a full analysis of EMDPs is rather lenghty and complicated, we leave this extension for future work.

Related Work. MDPs with mean payoff objectives (average reward criteria) have been heavily studied since the 60s (see, e.g., [27,31]). Several algorithms for computing optimal values and strategies have been developed for both finite-state systems (see e.g. [31,24,6,19]) as well as various types of infinite-state MDPs typically related to queueing systems (see, e.g., [29]). For an extensive survey see [31].

Markov decision processes with energy objectives have been studied in [7] as one-counter MDPs. Subsequently, several papers concerned MDPs with counters (resources) have been published (for a survey see [30], for recent work see e.g. [1]). A closely related paper [16] studies MDPs with combined energy-parity and mean-payoff-parity objectives (note, however, that the combination of energy with mean payoff is not studied in [16]).

A considerable amount of attention has been devoted to non-stochastic turn-based games with energy objectives [15,4]. Solving energy games belongs to $\mathsf{NP} \cap \mathsf{coNP}$ but no polynomial time algorithm is known. Energy games are polynomially equivalent to mean-payoff games [4]. Several papers are concerned with complexity of energy games (or equivalent problems, see e.g. [25,34,11,22]). For a more detailed account of results on energy games see [21]. Games with various combinations of objectives as well as multi-energy objectives have also been studied (see e.g. [32,2,10,28,18,16,5]), as well as energy constraints in automata settings [13].

Our work is closely related to the recent papers [12,23] where the combination of expected and worst-case mean-payoff objectives is considered. In particular, [23] considers a problem of optimizing the expected multi-dimensional mean-payoff under the condition that the mean-payoff in the first component is positive for all runs. At first glance, one may be tempted to "reduce" **[Q2]** and **[Q3]** to results of [23] as follows: Ask for a strategy which ensures that the mean-payoff in the first counter is non-negative for all runs, and then try to optimize the expected mean-payoff of the second counter. However, this approach does not work for several reasons. First, a strategy achieving non-negative mean-payoff in the first counter may still decrease the counter arbitrarily deep. So no matter what initial value of the counter is used, the zero counter value may be reached with positive probability. Second, the techniques developed in [23] do not work in the case of "balanced" EMDPs. Intuitively, balanced EMPDs are those where we inevitably need to employ strategies that balance the counter, i.e., the expected average change of the counter per transition is zero. In the framework of stochastic counter systems, the balanced subcase is often more difficult than the other subcases when the counters have a tendency to "drift" in some direction. In our case, the balanced EMDPs also require a special (and non-trivial) proof techniques based on martingales and some new "structural" observations. We believe that these tools can be adapted to handle the "balanced subcase" in even more general problems related to systems with more counters, MDPs over vector addition systems, and similar models.

2 Preliminaries

We use $\mathbb{Z}$, $\mathbb{N}$, $\mathbb{N}^+$, $\mathbb{Q}$, and $\mathbb{R}$ to denote the set of all integers, non-negative integers, positive integers, rational numbers, and real numbers, respectively. We assume familiarity with basic notions of probability theory, e.g., *probability space*, *random variable*, or the *expected value*. As usual, a *probability distribution* over a finite or countably infinite set A is a function $f : A \to [0,1]$ such that $\sum_{a \in A} f(a) = 1$. We call f *positive* if $f(a) > 0$ for each $a \in A$, *rational* if $f(a) \in \mathbb{Q}$ for each $a \in A$, and *Dirac* if $f(a) = 1$ for some $a \in A$.

Definition 1 (MDP). *A* Markov decision process (MDP) *is a tuple* $\mathcal{M} = (S, (S_\Box, S_\bigcirc), T, Prob, r)$, *where* S *is a finite set of* states, $(S_\Box, S_\bigcirc)$ *is a partitioning of* S

into the sets $S_\Box$ *of* controllable *states and* $S_\bigcirc$ *of* stochastic *states, respectively,* $T \subseteq S \times S$ *is a* transition relation, *Prob is a function assigning to every stochastic state* $s \in S_\bigcirc$ *a positive probability distribution over its outgoing transitions, and* $r\colon T \to \mathbb{Q}$ *is a* reward function. *We assume that* T *is* total, *i.e., for each* $s \in S$ *there is* $t \in S$ *such that* $(s,t) \in T$.

We use $Prob(s,t)$ as an abbreviation for $(Prob(s))(s,t)$, i.e., $Prob(s,t)$ is the probability of taking the transition (s,t) in s. For a state s we denote by $out(s)$ the set of transitions outgoing from s. A *finite path* is a sequence $w = s_0 s_1 \cdots s_n$ of states such that $(s_i, s_{i+1}) \in T$ for all $0 \leq i < n$. We write $len(w) = n$ for the length of the path. A *run* (or an *infinite path*) is an infinite sequence ω of states such that every finite prefix of ω is a finite path. For a finite path w, we denote by $Run_{\mathcal{M}}(w)$ the set of all runs having w as a prefix.

An *end component* of $\mathcal{M}$ is a pair (S', T'), where $S' \subseteq S$, $T' \subseteq T$, satisfying the following conditions: (1) for every $s \in S'$, we have that $out(s) \cap T' \neq \emptyset$; (2) if $s \in S' \cap S_\bigcirc$, then $out(s) \subseteq T'$; (3) the graph determined by (S', T') is strongly connected. Note that every end component of $\mathcal{M}$ can be seen as a strongly connected MDP (obtained by restricting the states and transitions of $\mathcal{M}$). A *maximal end component (MEC)* is an end component which is maximal w.r.t. pairwise inclusion. The MECs of a given MDP $\mathcal{M}$ are computable in polynomial time [20].

A *strategy* (or a *policy*) in an MDP $\mathcal{M}$ is a tuple $\sigma = (M, m_0, update, next)$ where M is a set of memory elements, $m_0 \in M$ is an initial memory element, $update : M \times S \to M$ a memory-update function, and $next$ is a function which to every pair $(s,m) \in S_\Box \times M$ assigns a probability distribution over $out(s)$. The function $update$ is extended to finite sequences of states in the natural way. We say that σ is *finite-memory* if M is finite, and *memoryless* if M is a singleton. Further, we say that σ is *deterministic* if $next(s,m)$ is Dirac for all $(s,m) \in S_\Box \times M$. Note that σ determines a function which to every finite path in $\mathcal{M}$ of the form ws, where $s \in S_\Box$, assigns the probability distribution $next(s,m)$, where $m = update(m_0, w)$. Slightly abusing our notion, we use σ to denote this function.

Fixing a strategy σ and an initial state s, we obtain the standard probability space $(Run_{\mathcal{M}}(s), \mathcal{F}, \mathbb{P}^\sigma_s)$ of all runs starting at s, where $\mathcal{F}$ is the σ-field generated by all *basic cylinders* $Run_{\mathcal{M}}(w)$, where w is a finite path starting at s, and $\mathbb{P}^\sigma_s\colon \mathcal{F} \to [0,1]$ is the unique probability measure such that for all finite paths $w = s_0 \cdots s_n$ it holds $\mathbb{P}^\sigma_s(Run_{\mathcal{M}}(w)) = \prod_{i=1}^{n} x_i$, where each x_i is either $\sigma(s_0 \cdots s_{i-1})(s_{i-1}, s_i)$, or $Prob(s_{i-1}, s_i)$, depending on whether s_{i-1} is controllable or stochastic (the empty product evaluates to 1). We denote by $\mathbb{E}^\sigma_s$ the expectation operator of this probability space.

We say that a run $\omega = s_0 s_1 \cdots$ is *compatible* with a strategy σ if $\sigma(s_0 \cdots s_i)(s_i, s_{i+1}) > 0$ for all $i \geq 0$ such that $s_i \in S_\Box$.

Definition 2 (EMDP). *An* energy MDP (EMDP) *is a tuple* $\mathcal{E} = (\mathcal{M}, E)$, *where* $\mathcal{M}$ *is a finite MDP and* E *is a function assigning to every transition an integer* update.

We implicitly extend all MDP-related notions to EMPDs, i.e., for $\mathcal{E} = (\mathcal{M}, E)$ we speak about runs and strategies in $\mathcal{E}$ rather than about runs and strategies in $\mathcal{M}$. A *configuration* of $\mathcal{E}$ is an element of $S \times \mathbb{Z}$ written as $s(n)$.

Given an EMDP $\mathcal{E} = (\mathcal{M}, E)$ and a configuration $s(n)$ of $\mathcal{E}$, we use $|\mathcal{E}|$ and $|s(n)|$ to denote the encoding size of $\mathcal{E}$ and $s(n)$, respectively, where the counter updates and rewards used in $\mathcal{E}$, as well as the n in $s(n)$, are written as (fractions of) binary numbers.

We also use $M_{\mathcal{E}}$ to denote the maximal non-negative integer u such that u or $-u$ is an update assigned by E to some transition.

Given a finite or infinite path $w = s_0 s_1 \cdots$ in $\mathcal{E}$ and an *initial configuration* $s_0(n_0)$, we define the *energy level* after i steps of w as $Lev^{(i)}_{n_0}(w) = n_0 + \sum_{i=0}^{i-1} E(s_i, s_{i+1})$ (the empty sum evaluates to zero). A configuration of $\mathcal{E}$ after i steps of w is then the configuration $s_i(n_i)$, where $n_i = Lev^{(i)}_{n_0}(w)$. Note that for all n and $i \geq 0$, $Lev^{(i)}_{n}$ can be understood as a random variable.

We say that a run ω initiated in s_0 is *safe* in a configuration $s_0(n_0)$ if $Lev^{(i)}_{n_0}(w) \geq 0$ for all $i \geq 0$. A strategy σ is safe in $s_0(n_0)$ if all runs compatible with σ are safe in $s_0(n_0)$. Finally, a configuration $s_0(n_0)$ is safe if there is at least one strategy safe in $s_0(n_0)$. The following lemma is straightforward.

Lemma 1. *If $s(n)$ is safe and $m \geq n$, then $s(m)$ is safe.*

To every run $\omega = s_0 s_1 \cdots$ in $\mathcal{E}$ we assign a mean payoff $MP(\omega)$ collected along ω defined as $MP(\omega) := \liminf_{n\to\infty}(\sum_{i=1}^{n} r(s_{i-1}, s_i))/n$. The function MP can be seen as a random variable, and for every strategy σ and initial state s we denote by $\mathbb{E}^{\sigma}_{s}[MP]$ its expected value (w.r.t. $\mathbb{P}^{\sigma}_{s}$).

Definition 3 (Energy-constrained value). *Let $\mathcal{E} = (\mathcal{M}, E)$ be an EMDP and $s(n)$ its configuration. The* energy-constrained mean-payoff value *(or simply the* value*) of $s(n)$ is defined by $Val(s(n)) := \sup\{\mathbb{E}^{\sigma}_{s}[MP] \mid \sigma$ is safe in $s(n)\}$. For every state s we also put $Val(s) := \lim_{n\to\infty} Val(s(n))$.*

Note that the value of every unsafe configuration is $-\infty$. We say that a strategy σ is *ε-optimal* in $s(n)$, where $\varepsilon \geq 0$, if σ is safe in $s(n)$ and $Val(s(n)) - \mathbb{E}^{\sigma}_{s}[MP] \leq \varepsilon$. A 0-optimal strategy is called *optimal*.

3 The Results

In this section we precisely formulate and prove the results about EMDPs announced in Section 1. Let $\mathcal{E} = (\mathcal{M}, E)$ be an EMDP. For every state s of $\mathcal{E}$, let *min-safe*(s) be the least $n \in \mathbb{N}$ such that $s(n)$ is a safe configuration. If there is no such n, we put *min-safe*$(s) = \infty$. The following lemma follows from the standard results on one-dimensional energy games [14].

Lemma 2. *There is a P^{EG} algorithm which computes, for a given EMDP $\mathcal{E} = (\mathcal{M}, E)$ and its state s, the value min-safe(s).*

Next, we present a precise definition of strongly connected and pumpable EMPDs. We say that $\mathcal{E}$ is *strongly connected* if for each pair of states s, t there is a finite path starting in s and ending in t. The pumpability condition is more specific.

Definition 4. *Let $\mathcal{E}$ be an EMDP and $s(n)$ a configuration of $\mathcal{E}$. We say that a strategy σ is* pumping in $s(n)$ *if σ is safe in $s(n)$ and $\mathbb{P}^{\sigma}_{s}(\sup_{i\geq 0} Lev^{(i)}_{n} = \infty) = 1$. Further, we say that $s(n)$ is* pumpable *if there is a strategy pumping in $s(n)$, and $\mathcal{E}$ is* pumpable *if every safe configuration of $\mathcal{E}$ is pumpable.*

The subclass of strongly connected pumpable EMDPs is denoted by SP-EMDP. Clearly, if $s(n)$ is pumpable, then every $s(m)$, where $m \geq n$, is also pumpable. Hence, for every $s \in S$, we define *min-pump*(s) as the least n such that $s(n)$ is pumpable. If there is no such n, we put *min-pump*$(s) = \infty$.

Intuitively, the condition of pumpability allows to increase the counter to an arbitrarily high value whenever we need. The next lemma says that we can compute a strategy which achieves that.

Lemma 3. *For every EMDP $\mathcal{E}$ there exist a memoryless* globally pumping *strategy σ, i.e. a strategy that is pumping in every pumpable configuration of $\mathcal{E}$. Further, there is a* P^{EG} *algorithm which computes the strategy σ and the value min-pump$(s) \leq 3 \cdot |S| \cdot M_{\mathcal{E}}$ for every state s of $\mathcal{E}$. The problem whether a given configuration of $\mathcal{E}$ is pumpable is* EG*-hard.*

Now we can state our results about SP-EMDPs.

Theorem 1. *For the subclass of SP-EMDPs, we have the following:*

1. *The problem whether a given EMDP $\mathcal{E}$ belongs to SP-EMDP is* EG*-hard and solvable by a* P^{EG} *algorithm.*
2. *The value of all safe configurations of a given SP-EMDP $\mathcal{E}$ is the same. Moreover, there is a* P^{EG} *algorithm which computes this value.*
3. *For every SP-EMDP $\mathcal{E}$ and every configuration $s(n)$ of $\mathcal{E}$, there is a strategy σ optimal in $s(n)$. In general, σ may require infinite memory, and there is a* P^{EG} *algorithm which computes a finite description of this strategy.*
4. *For every SP-EMDP $\mathcal{E}$, every configuration $s(n)$ of $\mathcal{E}$, and every $\varepsilon > 0$, there is a finite-memory strategy which is ε-optimal in $s(n)$. Further, there is a* P^{EG} *algorithm which computes a finite description of this strategy.*
5. *The gap threshold problem for SP-EMDPs is* EG *hard.*

In particular, note that ε-optimal strategies in SP-EMDPs require only finite memory (4.), but they are not easier to compute than optimal strategies (5.).

The following theorem summarizes the results for general EMDPs.

Theorem 2. *For general EMDPs, we have the following:*

1. *Optimal strategies may not exist in EMDPs that are either not strongly connected or not pumpable.*
2. *Given an EMDP $\mathcal{E}$, a configuration $s(n)$ of $\mathcal{E}$, and $\varepsilon > 0$, the value of $s(n)$ can be approximated up to the absolute error ε in time which is polynomial in $|\mathcal{E}|$, $|s(n)|$, $M_{\mathcal{E}}$, and $1/\varepsilon$.*
3. *Given an EMDP $\mathcal{E}$ and a state s of $\mathcal{E}$, the limit value Val(s) is computable in time polynomial in $|\mathcal{E}|$ and $M_{\mathcal{E}}$.*
4. *Let $\mathcal{E}$ be an EMDP, $s(n)$ a configuration of $\mathcal{E}$, and $\varepsilon > 0$. An ε-optimal strategy in $s(n)$ may require infinite memory. A finite description of a strategy σ which is ε-optimal strategy in $s(n)$ is computable in time polynomial in $|\mathcal{E}|$, $M_{\mathcal{E}}$, and $1/\varepsilon$.*
5. *The gap threshold problem for EMDPs is in* EXPTIME *and* PSPACE*-hard.*

Before proving Theorems 1 and 2, we introduce several tools that are useful for the analysis of strongly connected EMDPs. For the rest of this section, we fix a *strongly connected* EMDP $\mathcal{E} = (\mathcal{M}, E)$ where $\mathcal{M} = (S, (S_\Box, S_\bigcirc), T, Prob, r)$.

The key component for the analysis of $\mathcal{E}$ is the linear program $\mathcal{L}_\mathcal{E}$ shown in Figure 1 (left). The program is a modification of a program used in [6] for multi-objective mean-payoff optimization. For each transition e of $\mathcal{E}$ we have a non-negative variable f_e that intuitively represents the long-run frequency of traversals of e under some strategy (the fact that f_e's can be given this interpretation is ensured by the *flow constraints* introduced in the first three lines). The constraint on the fourth line then ensures that a strategy that visits each transition e with frequency f_e achieves a non-negative long-run change of the energy level. In other words, such a strategy ensures that the energy level does not have, on average, a tendency to decrease.

Intuitively, the optimal value of $\mathcal{L}_\mathcal{E}$ is the maximal expected mean payoff achievable under the constraint that the long-run average change (or *trend*) of the energy level is non-negative. Every safe strategy has to satisfy this constraint, because otherwise the probability of visiting a configuration with negative counter would be positive. Thus, using the methods adopted from [6], we get the following.

Lemma 4. *If there is a strategy σ that is safe in some configuration $s(n)$ of $\mathcal{E}$, then the linear program $\mathcal{L}_\mathcal{E}$ has a solution whose objective value is at least $\mathbb{E}^\sigma_s[MP]$.*

$$\textbf{maximize} \quad \sum_{e\in T} f_e \cdot r(e) \quad \text{subject to}$$

$$\sum_{e\in T} f_e = 1$$

$$\forall s \in S_\Box\colon \sum_{(s',s)\in T} f_{(s',s)} = \sum_{(s,s'')\in T} f_{(s,s'')}$$

$$\forall s \in S_\bigcirc,\ \forall (s,r) \in T\colon\ f_{(s,r)} = Prob(s,r) \cdot \sum_{(s',s)\in T} f_{(s',s)}$$

$$\sum_{e\in T} f_e \cdot E(e) \geq 0$$

$$\forall e \in T\colon\ f_e \geq 0$$

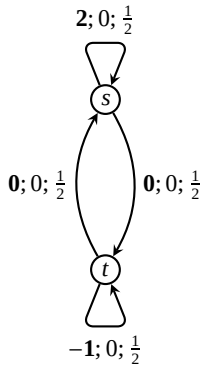

Fig. 1. A linear program $\mathcal{L}_\mathcal{E}$ with non-negative variables f_e, $e \in T$ (left), and an EMDP where the strategy corresponding to the solution of $\mathcal{L}_\mathcal{E}$ is not safe (right).

On the other hand, even if a strategy achieves a non-negative (or even positive) counter trend, it can still be unsafe in all configurations of $\mathcal{E}$. To see this, consider the EMDP of Figure 1 (right). There is only one strategy (the empty function), and it is easy to verify that assigning $1/4$ to each variable in $\mathcal{L}_\mathcal{E}$ solves the linear program with objective value $1/2$. However, for every m there is a positive probability that the decrementing loop on s is taken at least m times, and thus the strategy is not safe.

Although the program $\mathcal{L}_{\mathcal{E}}$ cannot be directly used to obtain a safe strategy optimizing the mean payoff, it is still useful for obtaining certain "building blocks" of such a strategy. To this end, we introduce additional terminology.

Let $\boldsymbol{f} = (f_e)_{e \in T}$ be an optimal solution of $\mathcal{L}_{\mathcal{E}}$, and let f^* be the corresponding optimal value of the objective function. A *flow graph* of $\boldsymbol{f}$ is a digraph $G_{\boldsymbol{f}}$ whose vertices are the states of $\mathcal{E}$, and there is an edge (s,t) in $G_{\boldsymbol{f}}$ iff there is a transition $e = (s,t)$ with $f_e > 0$. A *component* of $\boldsymbol{f}$ is a maximal set C of states that forms a strongly connected subgraph of $G_{\boldsymbol{f}}$. The set T_C consists of all $(s,t) \in T$ such that $s \in C$ and $f_{(s,t)} > 0$. A *frequency* of a component C is the number $f_C = \sum_{e \in T_C} f_e$. Finally, a *trend* and *mean-payoff* of a component C are the numbers $\mathit{trend}_C = \sum_{e \in T_C} (f_e / f_C) \cdot E(e)$ and $\mathit{mp}_C = \sum_{e \in T_C} (f_e / f_C) \cdot r(e)$.

Intuitively, the components of $\boldsymbol{f}$ are those families of states that are visited infinitely often by a certain strategy that maximizes the mean payoff while ensuring that the counter trend is non-negative. We show that our analysis can be simplified by considering only certain components of $\boldsymbol{f}$. We define a *type I core* and *type II core* of $\boldsymbol{f}$ as follows:

- A type I core of $\boldsymbol{f}$ is a component C of $\boldsymbol{f}$ such that $\mathit{trend}_C > 0$ and $\mathit{mp}_C \geq f^*$.
- A type II core of $\boldsymbol{f}$ is a pair C_1, C_2 of its components such that $\mathit{trend}_{C_1} \geq 0$, $\mathit{trend}_{C_2} \leq 0$, $f_{C_1} \cdot \mathit{trend}_{C_1} + f_{C_2} \cdot \mathit{trend}_{C_2} \geq 0$ and $f_{C_1} \cdot \mathit{mp}_{C_1} + f_{C_2} \cdot \mathit{mp}_{C_2} \geq f^*$.

The following lemma is easy.

Lemma 5. *Each optimal solution $\boldsymbol{f}$ of $\mathcal{L}_{\mathcal{E}}$ has a type I or a type II core. Moreover, a core of $\boldsymbol{f}$ (of some type) can be found in polynomial time.*

3.1 Strongly Connected and Pumpable EMDPs

In this subsection, we continue our analysis under the assumption that the considered EMPD $\mathcal{E}$ is not only strongly connected but also pumpable. Let $\boldsymbol{f}$ be an optimal solution to $\mathcal{L}_{\mathcal{E}}$ with optimal value f^*. We show how to use $\boldsymbol{f}$ and its core to construct a strategy optimal in every configuration $s(n)$ of $\mathcal{E}$. To some degree, the construction depends on the type of the core we use.

We start with the easier case when we compute a type I core C of $\boldsymbol{f}$. Consider two memoryless strategies: First, a memoryless deterministic globally pumping strategy π which is guaranteed to exist by Lemma 3. Second, we define a memoryless randomized strategy μ_C such that $\mu_C(s)(e) = f_e / f_C$ for all $s \in C$ and $e \in \mathit{out}(s)$, and $\mu_C(s)(e) = \kappa(s)(e)$ for all $s \notin C$ and $e \in \mathit{out}(s)$, where κ is a memoryless deterministic strategy in $\mathcal{E}$ ensuring that a state of T is reached with probability 1 (such a strategy exists as $\mathcal{E}$ is strongly connected). In order to combine these two strategies, we define a function low_n which assigns to a finite path w a value 1 if and only if there is $0 \leq j \leq \mathit{len}(w)$ such that $\mathit{Lev}_n^{(j)}(w) \leq L := M_{\mathcal{E}} + \max_{s \in S} \mathit{min\text{-}pump}(s)$ and $\mathit{Lev}_n^{(i)}(w) \leq H := L + |S| + 2|S|^2 \cdot M_{\mathcal{E}}$ for all $j \leq i \leq \mathit{len}(w)$; otherwise, $\mathit{low}_n(w) = 0$. We then define a strategy σ_n^* as follows:

$$\sigma_n^*(w)(e) = \begin{cases} \mu_C(\mathit{last}(w))(e) & \text{if } \mathit{low}_n(w) = 0 \\ \pi(\mathit{last}(w))(e) & \text{if } \mathit{low}_n(w) = 1. \end{cases}$$

Proposition 1. *Let $s(n)$ be a configuration of $\mathcal{E}$. Then σ^*_n is optimal in $s(n)$.*

Let us summarize the intuition behind the proof of Proposition 1. If the counter value is sufficiently high, we play the strategy μ prescribed by $\mathcal{L}_\mathcal{E}$ (i.e., we strive to achieve the mean payoff value f^*) until the counter becomes "dangerously low", in which case we switch to a pumping strategy that increases the counter to a sufficiently high value, where we again switch to μ. The positive counter trend achieved by μ ensures that if we start with a sufficiently high counter value, the probability of the counter *never* decreasing to dangerous levels is bounded away from zero. Moreover, once we switch to the pumping strategy π, with probability 1 we again pump the counter above $|S| \cdot H$ and thus switch back to μ. Hence, with probability 1 we eventually switch to strategy μ and use this strategy forever, and thus achieve mean payoff f^*.

Let us now consider the case where we compute a type II core of $\boldsymbol{f}$. The overall idea is similar as in the type I case. We try to execute a strategy that has non-negative counter trend and achieves the value f^* computed by $\mathcal{L}_\mathcal{E}$. This amounts to periodical switching between components C_1 and C_2, in such a way that the ratio of time spent in C_i tends to f_{C_i}. As in [6], this is done by fixing a large number N and fragmenting the play into infinitely many iterations: in the k-th iteration, we spend roughly $k \cdot N \cdot f_{C_1}$ steps in C_1, then move to C_2 and spent $k \cdot N \cdot f_{C_2}$ steps in C_2, then move back to C_1 and initialize the $(k{+}1)$-th iteration. Inside the component C_i we use the strategy μ_{C_i} defined above, until it either is time to switch to C_{3-i} or the counter becomes dangerously low. If the latter event happens, we immediately end the current iteration, switch to a pumping strategy, wait until a counter increases to a sufficient height, and then begin the $(k{+}1)$-th iteration. However, as the trend of μ_{C_2} is negative, the energy level tends to return to the value to which we increase the level during the pumping phase: it is thus no longer possible to prove, that we eventually stop hitting dangerously low levels. To overcome this problem, we use *progressive pumping*: the height to which we want to increase the counter after the "pumping mode" is switched on in the k-th iteration must increase with k, and it must increase asymptotically faster than $\sqrt{k}$. If this technical requirement is satisfied, we can use martingale techniques to show that progressive pumping decreases, with each iteration, the probability of drops towards dangerous levels. However, it also lengthens the time spent on pumping once such a period is initiated. To ensure that the fraction of time spent on pumping still tends to zero, we have to ensure that the threshold to which we pump increases *sublinearly* in k. In our proof we set the bound to roughly $k^{\frac{3}{4}}$ in order to satisfy both of the aforementioned constraints. More details in the appendix.

Proposition 2. *Each type II core of $\boldsymbol{f}$ yields a strategy optimal in $s(n)$.*

3.2 General EMDPs

In this section we prove Theorem 2. The two counterexamples required to prove part (1.) of the theorem are given in Fig. 2. On the left, there is a strongly connected but not pumpable EMDP (note that $t(0)$ is safe but not pumpable) where $Val(s(0)) = 5$, but there is no optimal strategy, and *every* strategy achieving a positive mean-payoff requires infinite memory (hence, this example also demonstrates that ε-optimal strategies

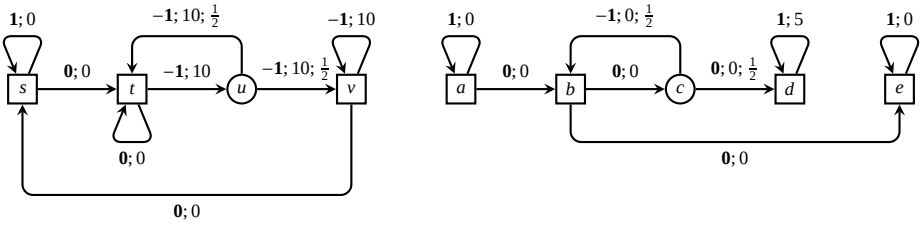

Fig. 2. Examples of EMDPs where optimal strategies do not exist in some configurations. Each transition is labeled by the associated counter update (in boldface), reward, and probability (only for the stochastic states u and c).

may require infinite memory, as stated in part (4) of Theorem 2). This is because the counter must be pumped to *linearly* larger and larger values when revisiting s to avoid reaching the configuration $t(0)$ with probability one (note that the probability of visiting $t(0)$ from $t(N)$ when using the transition (t, u) decays *exponentially* in N), yet ensuring that the mean payoff is equal to 5. Also note that if the counter was pumped to *exponentially* larger and larger values when revisiting s, the defining lim inf of mean payoff would be zero. On the right, there is pumpable but not strongly connected EMDP where $Val(a(0)) = 5$, but no optimal strategy exists in $a(0)$.

For the rest of this section, we fix an EMDP $\mathcal{E} = (\mathcal{M}, E)$. For simplicity, we assume that *for every $s \in S$ there is some $n \in \mathbb{N}$ such that the configuration $s(n)$ is safe*. The other control states can be easily recognized and eliminated (see Lemma 2).

Since $\mathcal{E}$ is not necessarily strongly connected, we start by identifying and constructing the MECs of $\mathcal{E}$ (this can be achieved in time polynomial in $|\mathcal{E}|$). Recall that each MEC of $\mathcal{E}$ can be seen as an EMDP, and each run eventually stays in some MEC [3]. Hence, we start by analyzing the individual MECs separately. Technically, we first assume that $\mathcal{E}$ is strongly connected.

The case when $\mathcal{E}$ is strongly connected. Consider a linear program $\mathcal{T}_{\mathcal{E}}$ which is the same as the program $\mathcal{L}_{\mathcal{E}}$ of Fig. 1 except for its objective function which is set to **maximize** $\sum_{t \in T} f_t \cdot E(t)$. In other words, $\mathcal{T}_{\mathcal{E}}$ tries to maximize the long-run average change of the energy level under the constraints given in $\mathcal{L}_{\mathcal{E}}$. Let $\boldsymbol{g} = (g_e)_{e \in T}$ be an optimal solution of $\mathcal{T}_{\mathcal{E}}$, and let g^* be the corresponding optimal value of the objective function. Now we distinguish two cases, which require completely different proof techniques.

Case A. $g^* > 0$.
Case B. $g^* = 0$.

We start with **Case A**. Note that if $g^* > 0$, then there exists a component D of $\boldsymbol{g}$ such that $trend_D \geq g^* > 0$. We proceed by solving the linear program $\mathcal{L}_{\mathcal{E}}$ of Fig. 1, and identifying the core of an optimal solution $\boldsymbol{f}$ of $\mathcal{L}_{\mathcal{E}}$. Recall that $\boldsymbol{f}$ can have either a type I core C, or a type II core C_1, C_2. In the first case, we set $E_1 := C$ and $E_2 := C$, and in the latter case we set $E_1 := C_1$ and $E_2 := C_2$. Let us fix some $\varepsilon > 0$. We compute positive rationals α_1, α_2 such

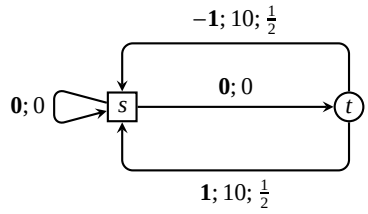

Fig. 3. An EMDP where the solution of $\mathcal{L}_\mathcal{E}$ is irrelevant.

- $\alpha_1 + \alpha_2 = 1$
- $\alpha_1 \cdot mp_{E_1} + \alpha_2 \cdot mp_{E_2} \geq f^* - \varepsilon/2$
- $\alpha_1 \cdot trend_{E_1} + \alpha_2 \cdot trend_{E_2} > 0$.

Observe that we can compute α_1, α_2 so that the length of the binary encoding of all of the above numbers is polynomial in $|\mathcal{E}|$ and $|\varepsilon|$. Now we construct a strategy which is safe and ε-optimal in every configuration with a sufficiently high counter value. Intuitively, we again just combine the two memoryless randomized strategies extracted from $\boldsymbol{f}$ (and possibly $\boldsymbol{g}$) in the ratio given by α_1 and α_2. Since the counter now has a tendency to increase under such a strategy, the probability of visiting a "dangerously low" counter value can be made arbitrarily small by starting sufficiently high (exponential height is sufficient for the probability to be smaller than ε). Hence, when such a dangerous situation occurs, we can permanently switch to *any* safe strategy (this is where our approach bears resemblance to [23]). For the finitely many configurations where the counter height is not "sufficiently large," the ε-optimal strategy can be computed by encoding these configurations into a finite MDP and optimizing mean-payoff in this MDP using standard methods.

Now consider **Case B**. If $g^* = 0$, the solution of $\mathcal{L}_\mathcal{E}$ is irrelevant, and we need to proceed in a completely different way. To illustrate this, consider the simple EMDP of Fig. 3. Here, the optimal solution $\boldsymbol{f}$ of $\mathcal{L}_\mathcal{E}$ produces $f^* = 5$ and assigns 1 to the transition (s, t). Clearly, we have that $Val(s(n)) = 0$ for an arbitrarily large n, so we cannot aim at approaching f^*. Instead, we show that if $g^* = 0$, then almost all runs produced by a safe strategy are *stable* in the following sense. We say that $s \in S$ is *stable at* $k \in \mathbb{Z}$ in a run $\omega = s_0 s_1 \cdots$ if there exists $i \in \mathbb{N}$ such that for every $j \geq i$ we have that $s_j = s$ implies $Lev_0^{(j)} = k$. Further, we say that s is *stable* in ω if s is stable at k in ω for some k. Note that the initial value of the counter does not influence the (in)stability of s in ω. Intuitively, s is stable in ω if it is visited finitely often, or it is visited infinitely often but from some point on, the energy level is the same in each visit. We say that a *run* is stable if each control state is stable in the run.

The next proposition represents another key insight into the structure of EMDPs. The proof is non-trivial and can be found in Appendix A.2.

Proposition 3. *Suppose that* $g^* = 0$, *and let* σ *be a strategy which is safe in* $s(n)$. *Then*

$$\mathbb{P}_s^\sigma(\{\omega \in Run(s) \mid \omega \text{ is stable }\}) = 1\,.$$

Due to Proposition 3, we can analyze the configurations of $\mathcal{E}$ in the following way. We construct a finite-state MDP where the states are the configurations of $\mathcal{E}$ with a non-negative counter value bounded by $|S| \cdot M_{\mathcal{E}}$. Transition attempting to decrease the counter below zero or increase the counter above $|S| \cdot M_{\mathcal{E}}$ lead to a special sink state with a self-loop whose reward is strictly smaller than the minimal reward used in $\mathcal{E}$. Then, we apply the standard polynomial-time algorithm for finite-state MDPs to compute the values in the constructed MDP, and identify a configuration $r(\ell)$ with the largest value. By applying Proposition 3, we obtain that $\mathit{Val}(t) = \mathit{Val}(r(\ell))$ for *every* $t \in S$. For every $\varepsilon > 0$, we can easily compute a bound $N_\varepsilon \in \mathbb{N}$ polynomial in $|\mathcal{E}|$, $M_{\mathcal{E}}$, and $1/\varepsilon$, and a memoryless strategy ϱ such that for every configuration $t(m)$ where $m \geq N_\varepsilon$ we have that the $\mathbb{P}_t^\varrho$ probability of all runs initiated in $t(m)$ that visit a configuration $r(k)$ for some $k \geq \ell$ without a prior visit to a configuration where the counter is "dangerously low" is at least $1 - (\varepsilon/R)$, where R is the difference between the maximal and the minimal transition reward in $\mathcal{E}$. Hence, a strategy which behaves like ϱ and "switches" either to a strategy which mimics the optimal behaviour in $r(\ell)$ (when a configuration $r(k)$ for some $k \geq \ell$ is visited) or to some safe strategy (when a configuration with dangerously low counter is visited) is ε-optimal in every configuration $t(m)$ where $m \geq N_\varepsilon$. For configurations with smaller counter value, an ε-optimal startegy can be computed by transforming the configurations with a non-negative counter value bounded by N_ε into a finite-state MDP and optimizing mean payoff in this finite-state MDP.

The case when $\mathcal{E}$ is not strongly connected. We finish by considering the general case when $\mathcal{E}$ is not strongly connected. Here, we again relay on standard methods for finite-state MDPs (see [31]). More precisely, we transform $\mathcal{E}$ into a finite-state MDP $\mathcal{M}[\mathcal{E}]$ in the following way. The states $\mathcal{M}[\mathcal{E}]$ consist of those states of $\mathcal{E}$ that do not appear in any MEC of $\mathcal{E}$, and for each MEC M of $\mathcal{E}$ we further add a fresh controllable state r_M to $\mathcal{M}[\mathcal{E}]$. The transitions of $\mathcal{M}[\mathcal{E}]$ are constructed as follows. For each r_M we add a self-loop whose reward is the limit value of the states of the MEC M in $\mathcal{E}$ (see the previous paragraph). Further, for every state s of $\mathcal{E}$, let $\hat{s}$ be either the state s of $\mathcal{M}[\mathcal{E}]$ or the state r_M of $\mathcal{M}[\mathcal{E}]$, depending on whether s belongs to some MEC M of $\mathcal{E}$ or not, respectively. For every transition (s,t) of $\mathcal{E}$ where s,t do *not* belong to the same MEC, we add a transition $(\hat{s},\hat{t})$ to $\mathcal{M}[\mathcal{E}]$. The rewards for all transitions, except for the self-loops on r_M, can be chosen arbitrarily.

Now we solve the standard mean-payoff optimization problem for $\mathcal{M}[\mathcal{E}]$, which can be achieved in polynomial time by constructing a suitable linear program [31]. The program also computes a *memoryless and deterministic* strategy σ which achieves the optimal mean-payoff $\mathit{MP}(s)$ in every state s of $\mathcal{M}[\mathcal{E}]$. Note that $\mathit{MP}(r_M)$ is *not* necessarily the same as the limit value of the states of M computed by considering M as a "standalone EMDP", because some other MEC with a better mean payoff can be reachable from M. However, the strategy σ eventually "stays" in some target r_M almost surely, and the probability of executing a path of length k before reaching a target r_M decays exponentially in k. Hence, for every $\delta > 0$, one can compute a bound L_δ such that the probability of reaching a target r_M in at most L_δ steps is at least $1-\delta$. Moreover, L_δ is polynomial in $|\mathcal{E}|$ and $1/\delta$.

Now we show that $\mathit{MP}(s) = \mathit{Val}(t)$ for every state t of $\mathcal{E}$ where $\hat{t} = s$. Further, we show that for every $\varepsilon \geq 0$, we can compute a sufficiently large $N_\varepsilon \in \mathbb{N}$ (still polynomial

in $|\mathcal{E}|$, $M_{\mathcal{E}}$, and $1/\varepsilon$) and a strategy ϱ such that for every initial configuration $t(m)$, where $m \geq N_\varepsilon$, we have that ϱ is safe in $t(m)$ and $\mathbb{E}_t^\varrho[MP] \geq MP(s) - \varepsilon$, where $\hat{t} = s$. The strategy ϱ "mimics" the strategy σ and eventually switches to some other strategy (temporarily or forever) in the following way:

- Whenever a configuration with a "dangerously low" counter value is encountered, ϱ switches to a safe strategy permanently.
- In a controllable state t of $\mathcal{M}$ which does not belong to any MEC of $\mathcal{E}$, ϱ selects a transition (t, u) such that $(t, \hat{u})$ is the transition selected by σ. In particular, if σ selects a transition (t, r_M), then ϱ selects a transition leading from t to some state of M.
- In a controllable state t of a MEC M, ϱ mimics σ in the following sense. If σ selects the transition (r_M, r_M), then ϱ permanently switches to the $\varepsilon/2$-optimal strategy for M constructed in the previous paragraph. If σ selects a different transition, then there must be a transition (s, t) of $\mathcal{E}$ where $s \in M$ such that $(r_M, \hat{t})$ is the transition selected by σ. Then ϱ temporarily switches to a strategy which strives to reach the control state s. When s is reached, ϱ restarts mimicking σ. Note that for every $\delta > 0$, one can compute a bound M_δ polynomial in $|\mathcal{E}|$ and $1/\delta$ such that the probability of reaching s in at most M_δ steps is at least $1 - \delta$.

We choose N_ε sufficiently large (with the help of the L_δ and M_δ introduced above) so that the probability of all runs initiated in $t(m)$, where $m \geq N_\varepsilon$, that reach a target MEC M with a counter value above the threshold computed for M and $\varepsilon/2$ by the methods of the previous paragraph, is at least $1 - \frac{\varepsilon}{2R}$, where R is the difference between the maximal and the minimal transition reward in $\mathcal{E}$. Hence, ϱ is ε-optimal in every $t(m)$ where $m \geq N_\varepsilon$. For configuration with smaller initial counter value, we compute an ε-optimal strategy as before.

Finally, let us note that Theorem 2 (5.) can be proven by reducing the following *cost problem* which is known to be PSPACE-hard [26]: Given an acyclic MDP $\mathcal{M} = (S, (S_\Box, S_\bigcirc), T, Prob, r)$, i.e., an MDP whose graph does not contain an oriented cycle, a non-negative cost function c (which assigns costs to transitions), an initial state s_0, a target state s_t, a probability threshold x, and a bound B, decide whether there is a strategy which with probability at least x visits s_t in such a way that the total cost accumulated along the path is at most B. The reduction is straightforward and hence omitted.

References

1. Abdulla, P., Ciobanu, R., Mayr, R., Sangnier, A., Sproston, J.: Qualitative analysis of vass-induced mdps. CoRR abs/1512.08824 (2015)
2. Abdulla, P., Mayr, R., Sangnier, A., Sproston, J.: Solving parity games on integer vectors. In: Proceedings of CONCUR 2013. Lecture Notes in Computer Science, vol. 8052, pp. 106–120. Springer (2013)
3. de Alfaro, L.: Formal verification of probabilistic systems. Phd. thesis, Stanford University, Stanford, CA, USA (1998)
4. Bouyer, P., Fahrenberg, U., Larsen, K., Markey, N., Srba, J.: Infinite runs in weighted timed automata with energy constraints. In: Proceedings of FORMATS 2008. LNCS, vol. 5215, pp. 33–47. Springer (2008)

5. Bouyer, P., Markey, N., Randour, M., Larsen, K.G., Laursen, S.: Average-energy games. In: Proceedings of GandALF 2015. pp. 1–15 (2015)
6. Brázdil, T., Brožek, V., Chatterjee, K., Forejt, V., Kučera, A.: Two views on multiple mean-payoff objectives in markov decision processes. Logical Methods in Computer Science 10(1) (2014)
7. Brázdil, T., Brozek, V., Etessami, K., Kučera, A., Wojtczak, D.: One-counter markov decision processes. In: Proceedings of SODA 2010. pp. 863–874. SIAM (2010)
8. Brázdil, T., Kiefer, S., Kučera, A.: Efficient analysis of probabilistic programs with an unbounded counter. J. ACM 61(6), 41:1–41:35 (Dec 2014), http://doi.acm.org/10.1145/2629599
9. Brázdil, T., Kiefer, S., Kučera, A., Novotný, P., Katoen, J.P.: Zero-reachability in probabilistic multi-counter automata. In: CSL-LICS'14. pp. 22:1–22:10. ACM (2014), http://doi.acm.org/10.1145/2603088.2603161
10. Brenguier, R., Cassez, F., Raskin, J.F.: Energy and mean-payoff timed games. In: Proceedings of the 17th International Conference on Hybrid Systems: Computation and Control. pp. 283–292. HSCC '14, ACM, New York, NY, USA (2014)
11. Brim, L., Chaloupka, J., Doyen, L., Gentilini, R., Raskin, J.: Faster algorithms for mean-payoff games. Formal Methods in System Design 38(2), 97–118 (2011)
12. Bruyère, V., Filiot, E., Randour, M., Raskin, J.F.: Meet Your Expectations With Guarantees: Beyond Worst-Case Synthesis in Quantitative Games. In: Mayr, E.W., Portier, N. (eds.) 31st International Symposium on Theoretical Aspects of Computer Science (STACS 2014). Leibniz International Proceedings in Informatics (LIPIcs), vol. 25, pp. 199–213. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany (2014)
13. Cachera, D., Fahrenberg, U., Legay, A.: An omega-Algebra for Real-Time Energy Problems. In: Proceedings of FSTTCS'15. LIPIcs, vol. 45, pp. 394–407. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany (2015)
14. Chakrabarti, A., de Alfaro, L., Henzinger, T.A., Stoelinga, M.: Resource Interfaces. In: Alur, R., Lee, I. (eds.) Proceedings of EMSOFT 2003. LNCS, vol. 2855, pp. 117–133. Springer, Heidelberg (2003)
15. Chakrabarti, A., de Alfaro, L., Henzinger, T., Stoelinga, M.: Resource interfaces. In: Proceedings of EMSOFT 2003. LNCS, vol. 2855, pp. 117–133. Springer (2003)
16. Chatterjee, K., Doyen, L.: Energy Parity Games. In: Abramsky, S., Gavoille, C., Kirchner, C., Meyer auf der Heide, F. (eds.) Proceedings of ICALP 2010, Part II. LNCS, vol. 6199, pp. 599–610. Springer Berlin Heidelberg (2010)
17. Chatterjee, K., Doyen, L.: Energy and Mean-Payoff Parity Markov Decision Processes. In: Proceedings of MFCS 2011. LNCS, vol. 6907, pp. 206–218. Springer (2011)
18. Chatterjee, K., Doyen, L., Henzinger, T., Raskin, J.: Generalized mean-payoff and energy games. In: Proceedings of FST&TCS 2010. LIPIcs, vol. 8, pp. 505–516. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik (2010)
19. Chatterjee, K., Komárková, Z., Křetínský, J.: Unifying two views on multiple mean-payoff objectives in Markov decision processes. In: Proceedings of LICS 2015. pp. 244–256 (2015)
20. Chatterjee, K., Henzinger, M.: Efficient and dynamic algorithms for alternating bÜchi games and maximal end-component decomposition. J. ACM 61(3), 15:1–15:40 (Jun 2014)
21. Chatterjee, K., Henzinger, M., Krinninger, S., Nanongkai, D.: Polynomial-time algorithms for energy games with special weight structures. Algorithmica 70(3), 457–492 (2014)
22. Chatterjee, K., Randour, M., Raskin, J.F.: Strategy synthesis for multi-dimensional quantitative objectives. Acta informatica 51(3-4), 129–163 (2014)
23. Clemente, L., Raskin, J.F.: Multidimensional beyond worst-case and almost-sure problems for mean-payoff objectives. In: Proceedings of LICS'15. pp. 257–268. IEEE Computer Society, Washington, DC, USA (2015)

24. Filar, J., Vrieze, K.: Competitive Markov Decision Processes. Springer-Verlag New York, Inc., New York, NY, USA (1996)
25. Gurvich, V., Karzanov, A., Khachiyan, L.: Cyclic games and an algorithm to find minimax cycle means in directed graphs. USSR Comput. Math. Math. Phys. 28(5), 85–91 (1990)
26. Haase, C., Kiefer, S.: The odds of staying on budget. In: Proceedings of ICALP 2015. Lecture Notes in Computer Science, vol. 9135, pp. 234–246. Springer (2015)
27. Howard, R.: Dynamic programming and Markov processes. The MIT press, New York London, Cambridge, MA (1960)
28. Juhl, L., Larsen, K.G., Raskin, J.: Optimal bounds for multiweighted and parametrised energy games. In: Liu, Z., Woodcock, J., Zhu, H. (eds.) Theories of Programming and Formal Methods - Essays Dedicated to Jifeng He on the Occasion of His 70th Birthday. Lecture Notes in Computer Science, vol. 8051, pp. 244–255. Springer (2013)
29. Kitaev, M., Rykov, V.: Controlled Queueing Systems. CRC Press (1995)
30. Kučera, A.: Playing games with counter automata. In: Proceedings of RP 2012. LNCS, vol. 7550, pp. 29–41. Springer (2012)
31. Puterman, M.L.: Markov Decision Processes. Wiley-Interscience (2005)
32. Velner, Y., Chatterjee, K., Doyen, L., Henzinger, T., Rabinovich, A., Raskin, J.: The complexity of multi-mean-payoff and multi-energy games. Information and Computation 241, 177–196 (2015)
33. Williams, D.: Probability with Martingales. Cambridge Mathematical Textbooks, Cambridge University Press, Cambridge, UK (1991)
34. Zwick, U., Paterson, M.: The complexity of mean payoff games on graphs. Theor. Comput. Sci. 158(1&2), 343–359 (1996)

Technical Appendix

A Proofs

In this section, we give full proofs that were omitted in the main body of the paper.

Lemma 3 *For every EMDP $\mathcal{E}$ there exists a memoryless strategy σ such that σ is pumping in every pumpable configuration of $\mathcal{E}$. Further, there is a P^{EG} algorithm which computes the strategy σ and the value min-pump$(s) \leq 3 \cdot |S| \cdot M_{\mathcal{E}}$ for every state s of $\mathcal{E}$. The problem whether a given configuration of $\mathcal{E}$ is pumpable is* EG*-hard.*

Proof. We reduce the problem of computing *min-pump* to the problem of computing minimal initial credit in *energy parity MDPs* [17], where we are required to find a safe strategy which visits with probability 1 a given set of states infinitely often. Given an EMDP $\mathcal{E}$ we construct a new EMDP $\mathcal{E}'$ by adding new states and transitions to $\mathcal{E}$. For each transition $e = (s, t)$ of $\mathcal{E}$ we add new controllable states s_e, s'_e and transitions $(s, s_e), (s_e, t)$, (s_e, s'_e), (s'_e, s_e) such that $E(s_e, s'_e) = -1$ and the other three transitions have energy update 0 (the reward of the new transitions is irrelevant). We require that some state if the form s'_e is visited infinitely often, i.e. that the counter is infinitely often decreased by 1. It is easy to verify that a configuration is pumpable if and only if it admits a safe strategy that satisfies this Büchi objective with probability one.

To determine minimal initial energy level needed to achieve the latter, in [17] the authors provide a polynomial reduction to determining the minimal initial level in energy Büchi games, a problem which is shown to be solvable by an P^{EG} algorithm in [16]. For memorylessness, assume that $\mathcal{E}$ is pumpable and let $\mathcal{E}''$ be an EMDP obtained by removing all transitions (s, t) such that *min-pump*$(s) + E(s, t) <$ *min-pump*(t), and removing all states s for which *min-pump*$(s) = \infty$. It is easy to check that *min-pump*-values of states in $\mathcal{E}''$ are the same as in $\mathcal{E}$, and moreover, *any* strategy in $\mathcal{E}''$ is safe in all safe configurations, so in particular there are no negative cycles in $\mathcal{E}''$. Moreover, in $\mathcal{E}''$, it must be possible to reach, from each state, a positive cycle with probability 1, otherwise the said state would be unpumpable with any initial energy level. Hence, we can pick a set Π of disjoint positive cycles such that at least one cycle in Π is reachable from each state of $\mathcal{E}''$ a define a memoryless strategy π in such a way that in a state on one of these cycles it selects a transition (of $\mathcal{E}''$) which keeps us on the cycle and in all other states it selects a transition which takes us closer to some of these cycles (optimal strategies for reachability are memoryless). It is then easy to show that π is a globally pumping strategy in $\mathcal{E}''$ and thus also in $\mathcal{E}$. □

A.1 Proofs of Section 3.1

Recall that we assume a fixed strongly connected and pumpable EMDP $\mathcal{E} = (\mathcal{M}, E)$ where $\mathcal{M} = (S, (S_{\Box}, S_{\bigcirc}), T, Prob, r)$. Let $\boldsymbol{f}$ be an optimal solution to the program $\mathcal{L}_{\mathcal{E}}$ of Figure 1 with optimal value f^*.

We start by considering the case where we compute a type I core of $\boldsymbol{f}$, i.e. on the proof of Proposition 1.

Proof of Proposition 1 Let C be a type I core of $\boldsymbol{f}$, $s(n)$ a configuration of $\mathcal{E}$, and let strategy σ^*_n be as in Proposition 1. If $s(n)$ is not safe, then σ^*_n any strategy is optimal in $s(n)$, so assume that $s(n)$ is safe. We prove that σ^*_n is optimal in $s(n)$. First note that σ^*_n is clearly safe in $s(n)$, since whenever we are configuration $t(\ell)$ with $\ell \leq M_{\mathcal{E}} + \mathit{min\text{-}pump}(t)$, the strategy μ_C starts to behave as a globally pumping strategy which never visits a configuration $t'(\ell')$ with $\ell' \leq \mathit{min\text{-}pump}(t')$, and moreover, such $t'(\ell')$ cannot be visited without previously visiting a configuration $t''(\ell'')$ with $\mathit{min\text{-}pump}(t'') \leq \ell'' \leq \mathit{min\text{-}pump}(t'') + M_{\mathcal{E}}$. So we focus on optimality of the mean payoff produced by σ^*_n,

First note that the memoryless strategy μ_C, one of the two constituent strategies of σ^*_n, achieves mean payoff f^* from each state of $\mathcal{E}$ [6, Lemma 4.3], and the long-run change of the energy level under μ_C is positive. In particular, it suffices to prove that with probability 1 the strategy σ^*_n eventually starts to behave as μ_C and sticks to this behaviour *forever*, or formally, that under σ^*_n it holds with probability one that for all but finitely many prefixes of w of the produced run we have $\mathit{low}_n(w) = 0$. To show this, we use the following fact:

Lemma 6. *The following holds for all $t \in S$ and $m \geq H$: For every state t, starting in configuration $t(m)$ with strategy μ_C, the probability that we eventually encounter a configuration $t'(m')$ with $m' \leq L$ is strictly smaller than* 1.

Proof. We first present the proof under the assumption that $C = S$ and $M_{\mathcal{E}} \leq 1$.

Since C has positive trend, the expected long-run change of the counter under μ_C is positive. From [9, Lemma 4] it follows that the probability of never hitting energy level $\leq L$ is positive for each initial energy level m greater than *some* finite bound $H' \geq L$. We prove that this finite bound can be assumed to be $L + |S| \leq H$.

For any $i \geq L+1$ denote by $\mathcal{Z}_i$ the set of all states s of $\mathcal{E}$ such that under strategy μ_C the probability of the energy level decreasing to L when starting in $s(i)$ equals 1. Note that $s \in \mathcal{Z}_i$ if and only if the following two conditions hold:

- When starting in $s(i)$ with strategy μ_C, the probability of decreasing the energy level to $i - 1$ is 1.
- Denoting by $\mathcal{R}_i$ the set of all states t such that configuration $t(i-1)$ is encountered with positive probability when starting in $s(i)$ with μ_C, it holds $\mathcal{R}_i \subseteq \mathcal{Z}_{i-1}$.

Note that if condition (1.) holds for at least one configuration of the form $s(i)$, it holds for all $s(i)$ s.t. $i \geq L$, since strategy μ_C is memoryless. As noted above, it holds for $s(H')$, so it holds for all $s(i)$ with $i \geq L$. Whether the second condition holds for $s(i)$ depends solely on $\mathcal{Z}_{i-1}$, as $\mathcal{R}_i = \mathcal{R}_{i'}$ for all i, i', again due to memorylessness of μ_C. Hence, if $\mathcal{Z}_i = \mathcal{Z}_{i+1}$, then $\mathcal{Z}_i = \mathcal{Z}_{i'}$ for all $i' \geq i$. Moreover, $\mathcal{Z}_i \supseteq \mathcal{Z}_{i+1}$ for all i, since if memoryless strategy μ_C almost surely decreases the energy level to L from some $u(i+1)$, it does the same from $u(i)$ as well. Hence, it must be the case that $\mathcal{Z}_{L+|S|} = \mathcal{Z}_{L+|S|+1}$ and thus $\mathcal{Z}_i = \mathcal{Z}_{H'}$ for all $i \geq L + |S|$. As shown above, $\mathcal{Z}_{H'} = \emptyset$, which finishes the proof for the special case.

Now we drop the assumption that $M_{\mathcal{E}} \leq 1$. We can then subdivide each transition (s,t) with $E(s,t) = e$ into a path of length $M_{\mathcal{E}}$ on which each edge is labelled by $e/M_{\mathcal{E}}$ (assignment of rewards is irrelevant). Thus, we reduce the proof to the case with

$M_{\mathcal{E}}$ at the cost of blowing-up the state space: the transformed EMDP $\mathcal{E}'$ has at most $|S|^2 \cdot M_{\mathcal{E}} + |S|$ states. The strategy μ_C can be straightforwardly carried over to this EMDP, and it is easy to check that the expected long-run change of the counter under μ_C is the same in $\mathcal{E}$ and $\mathcal{E}'$, in particular it is positive. Moreover, for each state t of the original MDP its *min-pump*-value is the same in both EMDPs. We can thus apply the results of the previous paragraph to $\mathcal{E}'$ and get that the probability of hitting energy level L from $s(i)$ using μ_C is less than 1 for each $i \geq L + 2|S|^2 \cdot M_{\mathcal{E}}$.

It remains to lift the assumption that $C = S$. So let $C \subset S$. Since μ_C reaches C almost surely from each state $s \in S$, and μ_C is memoryless, we know that from each such state s there is a path w of length at most $|S|$ such that w ends within C and is traversed with positive probability. So starting in configuration $s(L + |S| + 2|S|^2 \cdot M_{\mathcal{E}}) = s(H)$ and using strategy μ_C, we are guaranteed that with positive probability we hit a configuration $t(\ell)$ with $\ell \geq L + 2|S|^2 \cdot M_{\mathcal{E}}$ and $t \in C$ without hitting a configuration with energy level smaller than L. By previous paragraph, from $t(\ell)$ we have a positive probability of never going below L, which finishes the proof.

Now we finish the proof of Proposition 1. Suppose that with positive probability we infinitely often encounter the situation when the function low_n attains value 1. After each such occasion the strategy σ eventually switches back to behaving as μ_C, since π is a globally pumping strategy. When this switch occurs, there is a positive probability (bounded away from zero) that we will never encounter the situation with $low_n = 0$ again, as shown by the previous lemma. It follows, that the probability of infinitely often seeing such a situation is zero, a contradiction.

Proof of Proposition 2 To define an optimal strategy σ^*_n, we need additional notation: For $w = s_0 s_1 \cdots$ and $0 \leq i \leq len(w)$ we denote by $St(w, i)$ the state s_i.

We first prove a couple of useful general lemmas.

In the following we mean by "playing according to a memoryless strategy μ" that at each situation we select a distribution on actions prescribed by μ for the current state. We also use this terminology for history-dependent strategies: when saying that at some point (after observing a history w) we "play according to some strategy σ," we mean that from this point on, after seeing a history ww' we choose the distribution on actions given by $\sigma(w')$.

Lemma 7. *Let μ_1, μ_2 be memoryless strategies in $\mathcal{E}$, $p_1, p_2 \in [0, 1]$ numbers s.t. $p_1 + p_2 = 1$, $K \in \mathbb{N}$, $N \in \mathbb{N}$ the smallest number s.t. $p_1 \cdot N$ and $p_2 \cdot N$ are integers, and let q be any state of $\mathcal{E}$. Assume that both μ_1 and μ_2 determine a Markov chain with a single bottom strongly connected component (i.e. using μ_i, almost all runs have the same frequency of visits to a given state).*

For each $i \in \mathbb{N}$ let T_i be a probability distribution on $\mathbb{N}_0$ for which there exist a function $g : \mathbb{N} \to \mathbb{N}$ and a constant $c \in (0, 1)$ satisfying $\mathbb{P}(T_i \geq g(i)) \leq c^{-i}$ and $\lim_{i\to\infty} \sum_{i=1}^{n} g(i)/n^2 = 0$.

Finally, let σ be a strategy in $\mathcal{E}$ defined as follows: σ is played in stages. In stage $i \in \mathbb{N}$, we:

- *First play according to μ_1 for exactly $p_1 \cdot N \cdot i$ steps,*

– *then play according to μ_2 for exactly $p_2 \cdot N \cdot i$ steps,*
– *then play according to a memoryless deterministic strategy κ which guarantees reaching q with probability 1 (such a strategy exists due to $\mathcal{E}$ being strongly connected). We play according to κ until q is reached.*
– *Then, play according to a globally pumping strategy π (which is guaranteed to exist by Lemma 3). We play according to π for a random number of steps determined by a single draw from the distribution T_i.*
– *Then we proceed to stage $i+1$.*

Then for all states s it holds $\mathbb{E}_s^{\sigma}[MP] = p_1 \cdot \mathbb{E}_s^{\mu_1}[MP] + p_2 \cdot \mathbb{E}_s^{\mu_2}[MP]$.

Proof. Let us denote by $M_i^{\mu_1}$, $M_i^{\mu_2}$, M_i^{κ} and M_i^{π} the total rewards accumulated during the i-th stage playing according to μ_1, μ_2, κ and π. Denote by L_i^{κ} the number of steps made according to κ in the i-th stage. Slightly abusing notation, we use T_i to denote the number of steps made according to π in the i-th stage, and assume that $T_1, T_2, \ldots$ are independent. Denote by $\overline{L_i}$ the length of the i-the stage, i.e. $N \cdot i + L_i^{\kappa} + T_i$.

We use the following equation (which will be justified below): Almost surely,

$$MP = \lim_{n\to\infty} \frac{\sum_{i=1}^{n} M_i^{\mu_1} + M_i^{\mu_2} + M_i^{\kappa} + M_i^{\pi}}{\sum_{i=1}^{n} \overline{L_i}} = p_1 \cdot \mathbb{E}_s^{\mu_1}[MP] + p_2 \cdot \mathbb{E}_s^{\mu_2}[MP] \tag{1}$$

First, we show

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} M_i^{\mu_1} + M_i^{\mu_2} + M_i^{\kappa} + M_i^{\pi}}{\sum_{i=1}^{n} \overline{L_i}} = p_1 \cdot \mathbb{E}_s^{\mu_1}[MP] + p_2 \cdot \mathbb{E}_s^{\mu_2}[MP] \tag{2}$$

Then we finish the proof by proving (1). We have

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} M_i^{\mu_1} + M_i^{\mu_2} + M_i^{\kappa} + M_i^{\pi}}{\sum_{i=1}^{n} \overline{L_i}} = \tag{3}$$

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} M_i^{\mu_1} + M_i^{\mu_2}}{\sum_{i=1}^{n} N \cdot i} \frac{\sum_{i=1}^{n} N \cdot i}{\sum_{i=1}^{n} \overline{L_i}} + \lim_{n\to\infty} \frac{\sum_{i=1}^{n} M_i^{\kappa}}{\sum_{i=1}^{n} \overline{L_i}} + \lim_{n\to\infty} \frac{\sum_{i=1}^{n} M_i^{\pi}}{\sum_{i=1}^{n} \overline{L_i}} \tag{4}$$

assuming that the limits on the right-hand side exist.

One can easily show that, a.s.,

$$\begin{aligned}
\lim_{n\to\infty} \frac{\sum_{i=1}^{n} M_i^{\mu_1} + M_i^{\mu_2}}{\sum_{i=1}^{n} N \cdot i} &= \lim_{n\to\infty} \frac{\sum_{i=1}^{n} M_i^{\mu_1}}{\sum_{i=1}^{n} p_1 \cdot N \cdot i} \lim_{n\to\infty} \frac{\sum_{i=1}^{n} p_1 \cdot N \cdot i}{\sum_{i=1}^{n} N \cdot i} \\
&\quad + \lim_{n\to\infty} \frac{\sum_{i=1}^{n} M_i^{\mu_2}}{\sum_{i=1}^{n} p_2 \cdot N \cdot i} \lim_{n\to\infty} \frac{\sum_{i=1}^{n} p_2 \cdot N \cdot i}{\sum_{i=1}^{n} N \cdot i} \\
&= p_1 \cdot \mathbb{E}_s^{\mu_1}[MP] + p_2 \cdot \mathbb{E}_s^{\mu_2}[MP]
\end{aligned}$$

Here the last equality follows from the ergodic theorem for finite-state Markov chains (see e.g. [**?**]) applied to μ_1 and to μ_2.

So to prove (2) it suffices to prove the following equations (and apply (3)):

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} \overline{L_i}}{\sum_{i=1}^{n} N \cdot i} = 1 \tag{5}$$

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n M_i^\kappa}{\sum_{i=1}^n L_i} = 0 \tag{6}$$

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n M_i^\pi}{\sum_{i=1}^n \bar{L}_i} = 0 \tag{7}$$

We start by proving two auxiliary claims:

Claim (1).

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n L_i^\kappa}{n} < \infty$$

Proof (of the claim). let us define $L^\kappa_{i,s'}$ the number of steps played according to κ in the j-th stage where κ starts in s'. Given n denote by $n_{s'}$ the number of such stages up to the n-th stage. Then for every s' the $L^\kappa_{1,s'}, L^\kappa_{2,s'}, \dots$ are independent and identically distributed with $\mathbb{E}^\sigma_s(L^\kappa_{j,s'}) = \mathbb{E}^\sigma_s(L^\kappa_{1,s'}) < \infty$, and hence by invoking the strong law of large numbers for iid variables (see e.g. [33]) we obtain

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n L_i^\kappa}{n} = \lim_{n\to\infty} \frac{\sum_{s'} \sum_{j=1}^{n_{s'}} L^\kappa_{j,s'}}{n} = \sum_{s'} \lim_{n\to\infty} \frac{\sum_{j=1}^{n_{s'}} L^\kappa_{j,s'}}{n} =$$

$$\sum_{s'} \lim_{n\to\infty} \frac{\sum_{j=1}^{n_{s'}} L^\kappa_{j,s'}}{n_{s'}} \lim_{n\to\infty} \frac{n_{s'}}{n} \le \max_{s'} \lim_{n\to\infty} \frac{\sum_{j=1}^{n_{s'}} L^\kappa_{j,s'}}{n_{s'}} = \max_{s'} \mathbb{E}^\sigma_s(L^\kappa_{j,s'}) < \infty$$

This finishes the proof of Claim (1).

Claim (2).

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n T_i}{\sum_{i=1}^n i} = 0$$

Proof (of the claim). By our assumptions, $\mathbb{P}(T_i \ge g(i)) \le c^{-i}$ for all i and thus $\sum_{i=1}^\infty \mathbb{P}(T_i \ge g(i)) < \infty$. Hence, by Borel-Cantelli lemma (see [33]), for almost every run there is i' such that $T_i < g(i)$ for $i \ge i'$. However, then, a.s.,

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n T_i}{\sum_{i=1}^n i} = \lim_{n\to\infty} \frac{\sum_{i=i'}^n T_i}{\sum_{i=i'}^n i} < \lim_{n\to\infty} \frac{\sum_{i=i'}^n g(i)}{\sum_{i=i'}^n i} = \lim_{n\to\infty} \frac{\sum_{i=i'}^n g(i)}{\frac{n}{2}(n+1)} = 0$$

Here the last equality follows from our assumptions on g. This finishes the proof of the claim (2).

Let us prove the equation (5).

$$\begin{aligned}
\lim_{n\to\infty} \frac{\sum_{i=1}^n \bar{L}_i}{\sum_{i=1}^n N\cdot i} &= \lim_{n\to\infty} \frac{\sum_{i=1}^n N\cdot i + L_i^\kappa + T_i}{\sum_{i=1}^n N\cdot i} \\
&= \lim_{n\to\infty} \frac{\sum_{i=1}^n N\cdot i}{\sum_{i=1}^n N\cdot i} + \lim_{n\to\infty} \frac{\sum_{i=1}^n L_i^\kappa}{\sum_{i=1}^n N\cdot i} + \lim_{n\to\infty} \frac{\sum_{i=1}^n T_i}{\sum_{i=1}^n N\cdot i} \\
&= 1 + \lim_{n\to\infty} \frac{\sum_{i=1}^n L_i^\kappa}{n} \lim_{n\to\infty} \frac{n}{\sum_{i=1}^n N\cdot i} + \lim_{n\to\infty} \frac{\sum_{i=1}^n T_i}{\sum_{i=1}^n N\cdot i} \\
&= 1
\end{aligned}$$

The last equality follows from Claim (1) and Claim (2). This finishes the proof of (5).

Now let us prove (6):

$$\begin{aligned}\lim_{n\to\infty}\frac{\sum_{i=1}^{n} M_i^{\kappa}}{\sum_{i=1}^{n}\overline{L}_i} &\le \lim_{n\to\infty}\frac{\sum_{i=1}^{n} L_i^{\kappa}\cdot\max r}{\sum_{i=1}^{n}\overline{L}_i}\\ &= \max r\cdot\lim_{n\to\infty}\frac{\sum_{i=1}^{n} L_i^{\kappa}}{n}\cdot\lim_{n\to\infty}\frac{n}{\sum_{i=1}^{n} N\cdot i}\cdot\lim_{n\to\infty}\frac{\sum_{i=1}^{n} N\cdot i}{\sum_{i=1}^{n}\overline{L}_i}\\ &= 0\end{aligned}$$

Here the last equality follows from Claim (1) and the equation (5). Similarly, using Claim (2), we prove (7):

$$\begin{aligned}\lim_{n\to\infty}\frac{\sum_{i=1}^{n} M_i^{\pi}}{\sum_{i=1}^{n}\overline{L}_i} &\le \lim_{n\to\infty}\frac{\sum_{i=1}^{n} T_i\cdot\max r}{\sum_{i=1}^{n}\overline{L}_i}\\ &= \max r\cdot\lim_{n\to\infty}\frac{\sum_{i=1}^{n} T_i}{\sum_{i=1}^{n} N\cdot i}\cdot\lim_{n\to\infty}\frac{\sum_{i=1}^{n} N\cdot i}{\sum_{i=1}^{n}\overline{L}_i}\\ &= 0\end{aligned}$$

To finish the proof of Lemma 7 we prove that MP exists a.s. Then (1) follows from (2) and the fact that the sequence on the right-hand side of (1) is a subsequence of the mean-payoff defining sequence. Denote by MP_j the j-the average of the rewards obtained in the first j steps. Denote by k_j the number of stages completed in the first j steps.

Observe that

$$\frac{\sum_{i=1}^{k_j} M_i^{\mu_1}+M_i^{\mu_2}+M_i^{\kappa}+M_i^{\pi}}{\sum_{i=1}^{k_j}\overline{L}_i} \le MP_j \le \frac{\sum_{i=1}^{k_j} M_i^{\mu_1}+M_i^{\mu_2}+M_i^{\kappa}+M_i^{\pi}+\overline{L}_{k_j+1}\cdot\max r}{\sum_{i=1}^{k_j}\overline{L}_i}$$

Note that limits of the left-hand side and the right-hand side are equal as j goes to infinity, and of course, $\lim_{j\to\infty} MP_j = MP$. Indeed, observe

$$\begin{aligned}&\lim_{m\to\infty}\frac{\overline{L}_{m+1}}{\sum_{i=1}^{m}\overline{L}_i}\\ &\quad= \lim_{m\to\infty}\frac{N\cdot(m+1)+L_{m+1}^{\kappa}+T_{m+1}}{\sum_{i=1}^{m} N\cdot i+L_i^{\kappa}+T_i}\\ &\quad\le \lim_{m\to\infty}\frac{N\cdot(m+1)+L_{m+1}^{\kappa}+T_{m+1}}{\sum_{i=1}^{m} i}\\ &\quad\lim_{m\to\infty}\frac{N\cdot(m+1)+L_{m+1}^{\kappa}+T_{m+1}}{\sum_{i=1}^{m+1} i}\frac{\sum_{i=1}^{m+1} i}{\sum_{i=1}^{m} i}\\ &\quad= 0\end{aligned}$$

Here the last equality follows from Claim (1), Claim (2) and the fact that $\lim_{m\to\infty}\frac{\sum_{i=1}^{m+1} i}{\sum_{i=1}^{m} i} = 1$.

This finishes the proof of Lemma 7.

Now let C_1, C_2 be a type II core of $\boldsymbol{f}$, and $s(n)$ a configuration of $\mathcal{E}$. We again assume that $s(n)$ is safe.

As in the type I case, the components C_1, C_2 induces memoryless strategies μ_1, μ_2 such that for each $i \in \{1, 2\}$ the strategy μ_i behaves as follows: inside C_i it plays according to frequencies obtained from $\boldsymbol{f}$ and outside of C_i it behaves as a memoryless deterministic strategy for reaching C_i with probability 1. Note that both μ_i induce a Markov chain with a single bottom strongly connected component.

Let $p_1 = f_{C_1}$ and $p_2 = f_{C_2}$, $N \in \mathbb{N}$ the smallest number s.t. $p_1 \cdot N$ and $p_2 \cdot N$ are integers, and let q be an arbitrary state of $\mathcal{E}$. We define a strategy σ as follows: σ_1 is executed in stages. In stage $i \in \mathbb{N}$, we:

- First play according to μ_1 for exactly $p_1 \cdot N \cdot i$ steps,
- then play according to μ_2 for exactly $p_2 \cdot N \cdot i$ steps,
- then play according to a memoryless deterministic strategy κ which guarantees reaching q with probability 1 (such a strategy exists due to $\mathcal{E}$ being strongly connected). We play according to κ until q is reached.
- Then, play according to a globally pumping strategy π (which is guaranteed to exist by Lemma 3). We play according to π until the energy level is at least $TH + (i \cdot N)^{\frac{3}{4}}$, where $TH = \max_{q \in S} \mathit{min\text{-}pump}(q) + M_{\mathcal{E}}$.
- Then we proceed to stage $i + 1$.

Note that strategy σ *is not* safe in general.

Lemma 8. *Strategy σ_1 satisfies $\mathbb{E}_s^{\sigma}[MP] = p_1 \cdot \mathbb{E}_s^{\mu_1}[MP] + p_2 \cdot \mathbb{E}_s^{\mu_2}[MP]$. In particular, $\mathbb{E}_s^{\sigma}[MP] = f^*$.*

Proof. We use Lemma 7. The only thing we need to prove is to show that in each segment i, the random variable T_i denoting the time for which we play the globally pumping strategy π satisfies the condition in the assumptions of Lemma 7. That is, we need to find the right function g and constant c.

Note that in each stage we start playing according to π while in a state q. Memoryless strategy π induces a finite Markov chain M_π whose states are exactly the states of $\mathcal{E}$. Let $C_1, \ldots, C_\ell$ be all the bottom strongly connected components (BSCCs) of M_π that are reachable from q in M_π. It is easy to check that to satisfy the assumptions of Lemma 7 we need to prove the following:

- Denoting by T^1 the number of steps elapsed until one of the BSCCs $C_1, \ldots, C_\ell$ is reached, there exist a function $g_1 : \mathbb{N} \to \mathbb{N}$ and a constant $c_1 \in (0, 1)$ satisfying $\mathbb{P}_q^{\pi}(T^1 \geq g_1(i)) \leq c_1^{-i}$ and $\lim_{i \to \infty} \sum_{i=1}^{n} g_1(i)/n^2 = 0$ for all i.
- For all states t that belong to one of the components $C_1, \ldots, C_\ell$, there exist a function $g_2 : \mathbb{N} \to \mathbb{N}$ and a constant $c_2 \in (0, 1)$ satisfying $\mathbb{P}_t^{\pi}(T \geq g_2(i)) \leq c_2^{-i}$ and $\lim_{i \to \infty} \sum_{i=1}^{n} g_2(i)/n^2 = 0$ for all i.

The existence of g_1 and c_1 is easy, it follows, e.g. from [8, Lemma 5.1].

Now fix any state t as prescribed above. Note that from the construction of π it follows that it's counter trend trend_π from q (i.e. the number $\mathbb{E}_q^{\pi} \lim_{k \to \infty} \sum_{i=1}^{k} e_i(\omega)/k$, where $e_i(\omega)$ is the energy change on the i-th transition of ω) is positive (see the proof

of Lemma 3 – all cycles visited by the strategy have non-negative effect, and with probability 1 we infinitely often traverse a cycle of positive effect. Since π is memoryless, the probability of large gaps between two traversals of a positive cycle decays exponentially with the size of the gap, from which the result follows via standard computations). Since t is in a BSCC of the Markov chain induced by π, from [8] it follows that under π there is a *bounded-difference martingale*, a stochastic process $(\overline{m}^{(j)})_{j=0}^{\infty}$ given by $\overline{m}^{(j)}(\omega) = Lev_{\ell}^{(j)}(\omega) + \overline{z}(St(\omega, j)) - j \cdot trend_{\pi}$ for some weight function $\overline{z}\colon S \to \mathbb{R}$, where ℓ is the energy level in which we enter the BSCC in t.

Now any run ω initiated in t along which the energy level does not increase above $TH + (i \cdot N)^{\frac{3}{4}}$ in the first $W_i = (2 \cdot N \cdot i + TH)/trend_{\pi}$ steps satisfies $|\overline{m}^{(W_i)}(\omega) - \overline{m}^{(0)}(\omega)| \geq i - 2Z$, where $Z = \max_{t'} \overline{z}(t')$. From the Azuma's inequality [33] it follows that for all but finitely many i the probability $\mathbb{P}_t^{\pi}(T \geq W_i)$ is bounded from above by c_2^i for a suitable number $c_2 \in (0, 1)$. Hence, it suffices to put $g_2(i) = W_i$ for all such i. For the finitely many remaining i's we can set $g_2(i)$ to any number W such that the maximum among all these finitely many i's of the probability $\mathbb{P}_t^{\pi}(T \geq W)$ is smaller than, say $\frac{1}{2}$ (such a W exists, since π is pumping).

Now we modify σ to make it safe: in each stage, we play as prescribed above. However, if the current energy level falls below the threshold $TH = \max_{q \in S} min\text{-}pump(q) + M_{\mathcal{E}}$, we immediately skip to the second-to-last item, i.e. to the use of the globally pumping strategy π, which is played until the energy level surpasses the value prescribed for the current stage ($(i \cdot N)^{\frac{3}{4}}$). Denote this strategy σ_n^*. It is clear that σ_n^* is safe (it is actually pumping as well). It remains to prove that $\mathbb{E}_s^{\sigma_n^*}[MP] = \mathbb{E}_s^{\sigma}[MP]$, i.e. that σ_n^* is optimal.

We say that a stage i of σ_n^* *fails* if the energy level falls below TH during this stage. To prove that σ_n^* is optimal it suffices to prove that with probability 1, only finitely many stages of σ_n^* fail (and thus σ_n^* eventually starts to behave as σ forever). Due to Borel-Cantelli lemma it suffices to show that $\sum_{i=1}^{\infty} \mathbb{P}_s^{\sigma_n^*}(\sigma_n^* \text{ fails in stage } i) < \infty$. We prove that there is $i \in (0, 1)$ such that for all but finitely many i's the probability of failure in stage i is bounded by $c^{\sqrt{i}}$, which yields a converging infinite sum.

So let i be arbitrary and let t be an arbitrary state in which stage i starts. Note that stage i starts with energy level at least $TH + L_i$, where $L_i = (i \cdot N)^{\frac{3}{4}}$.

Consider the following events that may happen in stage i:

1. F_1: When starting in t, it takes at least $\frac{1}{6}L_i$ steps to reach C_1.
2. F_2: $\neg F_1$ and inside C_1 the counter increases by less than $f_{C_1} \cdot N \cdot i \cdot trend_{C_1} - \frac{2}{6}L_i$ before we start to play according to μ_2.
3. F_3: $\neg F_1$ and $\neg F_2$ and inside C_1 the counter decreases below TH before we start to play according to μ_2.
4. F_4: $\bigcap_{j=1}^{3} \neg F_j$ and upon starting to play according to μ_2, it takes at least $\frac{1}{6}L_i$ steps to reach C_2.
5. F_5: $\bigcap_{j=1}^{4} \neg F_j$ and inside C_2 the counter decreases by more than $f_{C_2} \cdot N \cdot i \cdot trend_{C_2} + \frac{1}{6}L_i$ before we start to play according to κ.
6. F_6: $\bigcap_{j=1}^{5} \neg F_j$ and upon starting to play according to κ, it takes at least $\frac{1}{6}L_i$ steps to reach q.

Note that if *none* of the events happens during the i-th stage, then this stage *does not* fail. Of particular interest here is the event F_5: note that if $\bigcap_{j=1}^{4} \neg F_j$ happens, then

when we enter C_2 while playing according to μ_2, our energy level is at least $L_i + f_{C_1} \cdot N \cdot i \cdot trend_{C_1} - \frac{4}{6}L_i$, so if F_5 holds, upon starting the play according to κ our energy level is at least $TH + L_i + f_{C_1} \cdot trend_{C_1} \cdot N \cdot i + f_{C_2} \cdot trend_{C_2} \cdot N \cdot i - \frac{5}{6}L_i = TH + \frac{1}{6}L_i$ (we have $f_{C_1} \cdot trend_{C_1} + f_{C_2} \cdot trend_{C_2} = 0$, since C_1, C_2 is a type II core of $\boldsymbol{f}$). Now to find c whose existence is postulated above, it is sufficient to find, for each of the above events, a number $d \in (0,1)$ such that for all but finitely many i's the probability of the said event is bounded by $d^{\sqrt{i}}$.

For events F_1, F_4, and F_6, we can again invoke Lemma 5.1. of [8]. The lemma proves that in a finite Markov chain (such as the one induced by a memoryless strategy for reaching some set of states) we can find a number $d' \in (0,1)$ such that the probability of not reaching a given almost-surely reachable set within ℓ steps is at most d'^{ℓ}. In our cases we have $\ell = i^{\frac{3}{4}} \cdot b$, where b is independent of i, which proves the existence of d.

For the remaining events we need to use arguments based on *martingales* [33]. Let us start with F_3. From Theorem 3.4. of [8] it follows that there is a *weight* function $z\colon S \to \mathbb{Q}$ such that for any $n \in \mathbb{Z}$ following stochastic process $(m^{(i)})_{i=0}^{\infty}$ is a martingale under μ_{C_1} when starting in C_1:[5]

$$m^{(i)}(\omega) = Lev_n^{(i)}(\omega) + z(St(\omega, i)) - i \cdot trend_{C_1}.$$

Moreover, from standard results on martingales, we get that if we denote by $\tau(\omega)$ the first point in time in which the energy level drops below TH, then the process $(\hat{m}^{(i)})_{i=0}^{\infty}$, where $\hat{m}^{(i)}(\omega) = m^{(\min\{i,\tau(\omega)\})}(\omega)$, is also a martingale. Moreover, both martingales have *bounded differences*, i.e. their one-step change is bounded uniformly over all runs and steps. Now any run ω initiated in some $u(\ell)$, $u \in C_1$, $\ell \geq TH + \frac{5}{6}L_i$ whose energy level drops below TH in the first $W = f_{C_1} \cdot N \cdot i$ steps[6] satisfies $|\hat{m}^{(W)}(\omega) - \hat{m}^{(0)}(\omega)| \geq \tau(\omega) \cdot trend_{C_1} + \frac{5}{6}L_i - 2Z \geq \frac{5}{6}L_i - 2Z$, where $Z = \max_{s \in S} |z(s)|$. The number on the right-hand side is positive for all but finitely many i. From the Azuma's inequality it follows that the probability of observing such a run is bounded by $d'^{(L_i - 2Z)^2/W} \leq d^{\sqrt{i}}$ for suitable numbers $d, d' \in (0,1)$ that are independent of i.

For event F_2 the argument is similar. Note that all runs in $\neg F_1$ make at least $f_{C_1} \cdot N \cdot i - \frac{1}{6}L_i$ steps inside C_1, since at most $\frac{1}{6}L_i$ steps were needed to reach C_1. If $\omega \in \neg F_1$ increases the counter by at least $f_{C_1} \cdot N \cdot k \cdot trend_{C_1} - \frac{1}{6}L_i$ during exactly $W' = f_{C_1} \cdot N \cdot i - \frac{1}{12}L_i$ steps, then it belongs to $\neg F_2$. So assume that $\omega \in \neg F_1$ increases the counter by at most $f_{C_1} \cdot N \cdot k \cdot trend_{C_1} - \frac{1}{6}L_i$ during exactly W' steps. Then $|m^{(W')} - m^{(0)}(\omega)| \geq \frac{1}{6}(i \cdot N)^{\frac{3}{4}} \cdot trend_{C_1} - 2Z$, where Z is as above. Again, this number is positive for all but finitely many i, and for all such i we can apply Azuma's inequality to get that probability of witnessing the small increase is at most $d^{\sqrt{i}}$, where d is a suitable number independent of k.

Event F_5 is handled in a way which is dual to F_2. We again use the construction from [8] to obtain a suitable martingale, which we analyse in almost the same way as in the previous paragraph. The only difference is that since F_2 has a negative trend, we now do not bound the probability of a small increase but that of a large decrease.

[5] Although [8] considers only a special case when $M_{\mathcal{E}} = 1$, the proof works also for our model without any modification.

[6] We can actually make smaller number of steps, because some steps might have been lost on reaching C_1. Nevertheless, overestimating the number of steps is sound.

A.2 Proofs of Section 3.2

Proposition 3 *Suppose that $g^* = 0$, and let σ be a strategy which is safe in $s(n)$. Then*

$$\mathbb{P}^{\sigma}_{s}(\{\omega \in \mathit{Run}(s) \mid \omega \textit{ is stable }\}) = 1\,.$$

Proof. We say that a run $\omega = s_0 s_1 \cdots$ in $\mathcal{E}$ is *drifting* if for every $k \in \mathbb{N}$ there exists $i \in \mathbb{N}$ such that for all $j \geq i$ we have that $\mathit{Lev}^{(j)}_0 \geq k$. Intuitively, a run is drifting if, for an arbitrary initial counter value, the energy level eventually stays above an arbitrarily large k along the run.

It follows from the results of [**?**] that the existence of a strategy π such that π is safe in some configuration $t(m)$ and $\mathbb{P}^{\pi}_{t}(\{\omega \in \mathit{Run}(t) \mid \omega \text{ is drifting }\}) > 0$ implies the existence of a positive solution of the program $\mathcal{T}_{\mathcal{E}}$.

Suppose that σ is a strategy safe in $s(n)$ such that

$$\mathbb{P}^{\sigma}_{s}(\{\omega \in \mathit{Run}(s) \mid \omega \text{ is stable }\}) < 1\,.$$

We show that there exist a configuration $t(m)$ and a strategy π with the above properties, and thus derive a contradiction. For every $q \in S$, all $A, B \subseteq S$ where $A \cap B = \emptyset$, and all $f : A \to \mathbb{Z}$, let $\mathit{Run}[A_f, B](q)$ be the set of all $\omega \in \mathit{Run}(q)$ such that the set of all control states that appear infinitely often along ω is precisely $A \cup B$, the set of all control states that are not stable in ω is precisely B, and every control state $r \in A$ is stable at $f(r)$ in ω. Clearly, there must be some A, f, B such that $B \neq \emptyset$ and $\mathbb{P}^{\sigma}_{s}(\mathit{Run}[A_f, B](s)) > 0$. For the rest of this proof, we fix such A, f, B.

For every configuration $r(\ell)$, we define the $[A_f, B]$-*value* of $r(\ell)$ as follows:

$$V_{[A_f,B]}(r(\ell)) := \sup\{\mathbb{P}^{\varrho}_{r}(\mathit{Run}[A_f, B](r)) \mid \varrho \text{ is safe in } r(\ell)\}.$$

Observe that $V_{[A_f,B]}(r(i)) \geq V_{[A_f,B]}(r(j))$ if $i \geq j$. We prove the following:

A. For every $r \in A$, let $r(\ell)$ be the configuration where $\ell = n + f(r)$. Then $V_{[A_f,B]}(r(\ell)) = 1$.
B. If $A \neq \emptyset$, then there is a configuration $r(\ell)$ such that $r \in B$ and $V_{[A_f,B]}(r(\ell)) = 1$.

To prove A., let us suppose that there is $r \in A$ such that $V_{[A_f,B]}(r(\ell)) = 1 - \delta$, where $\ell = n + f(r)$ and $\delta > 0$. Let $\omega \in \mathit{Run}[A_f, B](s)$, and consider the sequence of configurations visited by ω from the initial configuration $s(n)$. Since $r(\ell)$ appears infinitely often in this sequence, we obtain that $\mathbb{P}^{\sigma}_{s}(\mathit{Run}[A_f, B](s)) = 0$, which is a contradiction.

To prove B., suppose that there is some $q \in A$, but for all $r \in B$ and $\ell \in \mathbb{N}$ we have that $V_{[A_f,B]}(r(\ell)) < 1$. By A., we obtain $V_{[A_f,B]}(q(m)) = 1$ for a suitable m. For every $\omega \in \mathit{Run}[A_f, B](q)$, consider the sequence of configurations visited by ω from the initial configuration $s(m)$, and let $r(\ell)$ be the first configuration in this sequence such that $r \in B$. Clearly, $\ell \leq m + |S| \cdot M_{\mathcal{E}}$. Let

$$V = \max\{V_{[A_f,B]}(u(j)) \mid u \in B,\, j \leq m + |S| \cdot M_{\mathcal{E}}\}\,.$$

Since $V = 1 - \delta$ for some $\delta > 0$, for every strategy ϱ safe in $s(m)$ we obtain that $\mathbb{P}^{\varrho}_{q}(\mathit{Run}[A_f, B](q)) \leq 1 - \delta$, which contradicts $V_{[A_f,B]}(q(m)) = 1$.

The existence of π is now proved separately for each of the following two cases:

Case I. Suppose that $V_{[A_f,B]}(r(\ell)) = 1$ for some $r \in B$ and $\ell \in \mathbb{N}$. Let us further assume that ℓ is the *least* i such that $V_{[A_f,B]}(r(i)) = 1$. A finite path w from r to r of length j is *increasing* if $Lev_0^{(j)}(w) > 0$. We claim that for every $\varepsilon > 0$, there exist a strategy σ_ε safe in $r(\ell)$, and $N_\varepsilon \in \mathbb{N}$, such that the $\mathbb{P}_r^{\sigma_\varepsilon}$-probability of all runs initiated in r that start with an increasing path of length at most N_ε is at least $1 - \varepsilon$. Before proving this claim, let us show how it implies the existence of the promised $t(m)$ and π. The role of $t(m)$ is taken over by $r(\ell)$. The strategy π is constructed as follows. Let $\varepsilon_i = 8^{-i}$ for all $i \in \mathbb{N}^+$. Consider the strategies σ_{ε_i} and the bounds N_{ε_i} for all $i \in \mathbb{N}^+$. The strategy π is defined inductively as follows:

- At the starting state r, the strategy π "switches" to σ_{ε_1}.
- Whenever π "switches" to σ_{ε_j}, it starts to simulate the strategy σ_{ε_j}. If an increasing path is encountered in the first N_{ε_j} steps from the previous switch, then π immediately "switches" to $\sigma_{\varepsilon_{j+1}}$. Otherwise, π keeps simulating σ_{ε_j} forever.

It follows immediately from the construction of π that π is safe in $r(\ell)$ and the probability of all runs with infinitely many "switches" is at least $3/4$. Since all runs with infinitely many switches are drifting, we are done.

So, it remains to prove the above claim. Let us fix some $\varepsilon > 0$. Let $\kappa = \varepsilon\delta/2$, where δ is either 1 or $1 - V_{[A_f,B]}(r(\ell-1))$, depending on whether $\ell = 0$ or $\ell > 0$, respectively (note that $\delta > 0$). We put $\sigma_\varepsilon := \varrho$, where ϱ is a strategy safe in $r(\ell)$ such that $\mathbb{P}_r^\varrho(Run[A_f, B](r)) \geq 1 - \kappa$. Note that ϱ is guaranteed to exist, because the $[A_f, B]$-value of $r(\ell)$ is equal to one. Since $r \in B$, for *every* run $\omega = s_0 s_1 s_2 \cdots$ in $Run[A_f, B](r)$ there exist $i < j$ such that $s_i = s_j = r$ and $Lev_0^{(i)}(\omega) < Lev_0^{(j)}(\omega)$. We say that ω is *good* if there are $i < j$ with the above properties such that, in addition, for every $k \leq j$ we have that $s_k = r$ implies $Lev_0^{(k)}(\omega) \geq 0$. Now we check that

$$\mathbb{P}_r^\varrho(\{\omega \in Run[A_f, B](r) \mid \omega \text{ is good }) \geq 1 - \frac{\varepsilon}{2}\,.$$

If $\ell = 0$, the above inequality follows immediately, because then ϱ is safe in $r(0)$. If $\ell > 0$, then the $\mathbb{P}_r^\varrho$ probability of all $\omega \in Run(r)$ that are *not* good runs of $Run[A_f, B](r)$ cannot exceed $\varepsilon/2$, because otherwise, even if all of these runs belong to $Run[A_f, B](r)$, we obtain that $\mathbb{P}_r^\varrho(Run[A_f, B](r))$ is smaller than

$$(1 - \frac{\varepsilon}{2}) + \frac{\varepsilon}{2}(1 - \delta) \;=\; 1 - \kappa\,,$$

which is a contradiction. Since every good run of $Run[A_f, B](r)$ can be recognized after a finite prefix, there must by some N_ε such that the $\mathbb{P}_r^\varrho$ probability of all good runs of $Run[A_f, B](r)$, where the length this prefix is bounded by N_ε, is at least $1 - \varepsilon$.

Case II. Suppose that $V_{[A_f,B]}(r(\ell)) < 1$ for all $r \in B$ and $\ell \in \mathbb{N}$. Note that this implies $A = \emptyset$ by applying claim B. above. For every $\omega \in Run[A_f, B](s)$, let α_ω be the sequence of $[A_f, B]$-values of the configurations visited by ω from the initial configuration $s(n)$. Further, let $Lim[A_f, B](\omega) = \liminf_{n\to\infty} \alpha_\omega$. We claim that

$$\mathbb{P}_s^\sigma(\{\omega \in Run[A_f, B](s) \mid Lim[A_f, B](\omega) < 1) = 0\,.$$

Again, let us first show that this claim implies the existence of the promised $t(m)$ and π. In this case, the role of $t(m)$ is played by $s(n)$, and π is chosen as σ. Since almost

all $\omega \in \mathit{Run}[A_f, B](s)$ satisfy $\mathit{Lim}[A_f, B](\omega) = 1$, it suffices to show that every run $\omega = s_0 s_1 \cdots$ of $\mathit{Run}[A_f, B](s)$ such that $\mathit{Lim}[A_f, B](\omega) = 1$ is drifting. However, since $V_{[A_f,B]}(r(\ell)) < 1$ for all $r \in B$ and $\ell \in \mathbb{N}$, it follows immediately that for all $r \in B$ and $k \in \mathbb{N}$ there exists $i \in \mathbb{N}$ such that for all $j \geq i$ we have that $s_j = r$ implies $\mathit{Lev}_0^{(j)}(\omega) \geq k$. So, ω is indeed drifting.

It remains to prove the above claim. It suffices to show that for every fixed $\varepsilon > 0$ we have that

$$\mathbb{P}_s^{\sigma}(\{\omega \in \mathit{Run}[A_f, B](s) \mid \mathit{Lim}[A_f, B](\omega) < 1 - \varepsilon) = 0\,.$$

Let $\omega \in \mathit{Run}[A_f, B](s)$ be a run such that $\mathit{Lim}[A_f, B](\omega) < 1 - \varepsilon$, and let us consider the sequence of configurations visited by ω from the initial configuration $s(n)$. Clearly, this sequence visits infinitely often a configuration whose $[A_f, B]$-value is bounded by $1 - \varepsilon$, which implies that the total probability of all such runs is zero. □

A.3 A Proof of Theorem 1 (5.)

As explained in Section 1, the problem whether a given configuration of EMDP is safe is equivalent to solving the corresponding energy game (with the same transition structure as the EMDP). To finish the proof of Theorem 1 (5.), we need to show that it suffices to restrict to pumpable EMDPs.

So let us fix an EMDP $\mathcal{E} = (\mathcal{M}, E)$ where $\mathcal{M} = (S, (S_\Box, S_\bigcirc), T, \mathit{Prob}, r)$. We define an EMDP $\mathcal{E}' = (\mathcal{M}', E')$ where the set of states is $S \cup T$, from each $s \in S$ there are transitions to all elements of $\mathit{out}(s)$, from each $(s, s') \in T$ there are transitions to (s, s') and to s'. The set of stochastic states of $\mathcal{M}'$ is $S_\bigcirc \cup T$. The probability of each transition $(s, (s, s'))$, here $s \in S_\bigcirc$, in $\mathcal{M}'$ is equal to the probability of (s, s') in $\mathcal{M}$. The probability of each transition $((s, s'), (s, s'))$ in $\mathcal{M}'$ is equal to $\frac{1}{2}$. The energy update function E' is defined by $E'(s, (s, s')) = E(s, s')$ and $E'((s, s'), (s, s')) = \max_{e \in T} E(e) + 1$ and $E'((s, s'), s') = 0$. The reward function in $\mathcal{M}'$ can be defined arbitrarily (we are concerned only with safety).

Now note that a configuration $s(n)$ is safe in $\mathcal{M}$ iff $s(n)$ is safe in $\mathcal{M}'$. So $\mathit{Val}(s(n)) > -\infty$ in $\mathcal{M}'$ iff $\mathit{Val}(s(n)) > -\infty$ in $\mathcal{M}$ iff $s(n)$ is safe in the corresponding energy game on $\mathcal{M}$. Also, note that $\mathcal{M}'$ is pumpable since in every (s, s') the counter may be pumped above any bound with a positive probability, which eventually happens with probability one.