Strategy Complexity and Computational Complexity in Simple Stochastic Games with Multiple Objectives

Mohan Sai Teja Dantam

Doctor of Philosophy

Laboratory for Foundations of Computer Science
School of Informatics
University of Edinburgh

2026

Abstract

This thesis considers problems related to *multi-objective* settings in turn-based, zero-sum stochastic games on graphs with perfect information. When a player must pursue two or more objectives simultaneously, novel strategic behaviours emerge that are rarely observed in single-objective games. Typical single-objective methods either prove that pure stationary (memoryless deterministic, MD) strategies are sufficient, or invoke known strategy-lifting theorems to reduce the analysis to a simpler one-player game. By contrast, in multi-objective games MD strategies usually fail: the player must either store some history to shift focus between objectives, randomize between actions, or do both. The same difficulty arises already in single-player stochastic games (MDPs), thereby invalidating the hypotheses required by the lifting theorems. Finally, multiple objectives raise additional questions–most notably Pareto optimality, which captures the trade-off between the different components of a combined goal. We consider three different scenarios in this vast space of possibilities and provide solutions in these cases.

First, we study stochastic games $\mathcal{G}$ with the *Energy-Parity* objective, which combines a quantitative energy constraint with a qualitative parity condition. The Max tries to avoid running out of energy while simultaneously satisfying a parity condition. We present an algorithm to *approximate* the value of a given configuration in 2-NEXPTIME. Moreover, the corresponding ε-optimal strategies for either player use no more than $\mathcal{O}\big(\text{2-EXP}(\|\mathcal{G}\|) \cdot \log\big(\frac{1}{\varepsilon}\big)\big)$ memory modes.

Second, we analyse finite-state Markov decision processes equipped with the combined *Energy–Mean-Payoff* objective. The controller tries to avoid running out of energy while simultaneously attaining a strictly positive mean-payoff in a second reward dimension.

We establish that *finite memory* suffices for almost surely winning strategies for the Energy–Mean-Payoff objective. This contrasts with the Energy-Parity setting, where almost surely winning strategies generally need infinite memory.

We prove that exponential memory is sufficient (even for deterministic strategies) and necessary (even for randomized strategies) for almost surely winning Energy–Mean-Payoff. The same upper bound applies when the mean-payoff requirement is generalized to its multidimensional variant.

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

Third, we study the strategy complexity (*i.e.*, memory and randomization) of optimal strategies in stochastic games. For shift-invariant inverse-submixing objectives, it is known how to lift optimal finite-memory strategies from MDPs to games with an exponential increase in the number of memory modes. We demonstrate the corresponding lower bound, *i.e.*, the extra exponential memory is required in general, even if one allows randomization in both actions and updates.

While the worst case looks exponential, we show that it is easier for the conjunction of two well-studied objectives, namely the *positive Mean-Payoff–Parity* objective ($\texttt{MP} > 0 \cap \texttt{EPAR}$), which is also shift-invariant inverse-submixing. In (Maximizing) MDPs, it is known that optimal *deterministic* strategies require at least exponential memory. We prove that, with randomization, optimal strategies can be chosen memoryless. However, in stochastic games, while the lifting theorem provides an exponential upper bound, we prove that optimal *randomized* strategies require at most polynomial memory (equal to the number of even colors) and we give a matching family of games that proves the lower bound. Optimal *deterministic* strategies, on the other hand, need exponential memory.

Finally, we prove that an alternative lifting technique–one that works for memoryless (or, respectively, finite-memory) deterministic strategies–does not extend to memoryless randomized strategies.

Lay Summary

This thesis explores the mathematical strategies required for decision-making in complex environments, modeled as "games" played on graphs. These games represent scenarios where a controller must interact with an unpredictable or antagonistic environment to achieve a goal. In games with a single objective, simple reactive strategies are often sufficient. However, real-world systems usually need to balance multiple, conflicting goals simultaneously. We demonstrate that in these multi-objective settings, such simple strategies fail; to succeed, a system must typically remember a history of past events or make randomized choices to navigate the trade-offs between different goals.

The core of this research investigates the "complexity cost" of finding optimal strategies in these difficult scenarios. We analyze exactly how much memory and computational power are strictly necessary to guarantee success across various combinations of objectives. Our findings reveal a distinct trade-off: while some problems require the system to track a vast, exponentially growing amount of information, others become significantly simpler to solve if the controller is allowed to behave randomly rather than adhering to a rigid, deterministic plan. By establishing these mathematical boundaries, this work helps define the limits of efficient algorithm design for autonomous systems operating under complex, multi-layered constraints.

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my primary supervisor, Richard Mayr. Without his guidance, this thesis would not have been possible. Richard introduced me to the field of game theory and research on games on graphs, and I will be forever grateful for his time and this wonderful opportunity.

I would also like to thank my second supervisor, Kousha Etessami, for his unwavering availability and advice. My discussions with him, whether about research technicalities or general life ambitions, were always honest and incredibly helpful. I only wish I had taken the opportunity to tap into his vast knowledge even more during my PhD.

I extend my thanks to Rik Sarkar for giving me the chance to be involved in his research in Machine Learning, and to my examiners, Aris Filos-Ratsikas and Mickael Randour, whose insightful comments and suggestions undoubtedly improved this thesis.

I am grateful to the University of Edinburgh and the LFCS for funding my PhD and conference travels. A special thanks to Patrick, Jonathan, and the entire IGS team for handling the many administrative questions I encountered and for organizing and funding some memorable social events.

I would not have been able to embark on a PhD were it not for the professors and teachers I had the pleasure of learning from, from my school years through to my Masters at ENS Paris-Saclay. I am especially thankful to S. Akshay and Amaury Pouly for their patience and guidance when I was taking my first steps as a researcher. I also thank Roopsha Samanta, Joël Ouaknine, Markus Whiteland, and Engel Lefaucheux for the opportunity to learn so much about research during my time with them.

My time in Edinburgh was enriched by a wonderful academic community. Thanks to Sal, Wenyue, Shuai, Sunglin, Christopher, Zeyu, Georgios, and Yen for being fantastic office mates. Thank you to Eric, Asif, Tobias, Nana, Karim, and many others for the engaging discussions over the years. I also thank Georgios and Jakub for introducing me to the Game Theory seminar, and Charalampos for organizing it.

I am grateful to Tim, Oleg, Chak, and the entire TSRL team at J.P. Morgan for giving me the chance to be part of something amazing during my internship.

My time in London was made all the more enjoyable courtesy of the amazing company I had—thanks to Santosh, Sanskaar, Anjali, Uday, Mano, Surabhi, and Divyank.

Beyond research, I had the opportunity to meet many people who made this journey special. Thanks to Saurabh, Ruchika, Arjun, Nicolas, Hanyu, Stavros, Khalis, Tim, Luis, and Keshav for making my three years at Pentland House very enjoyable. Away from studies, I was glad to be a part of the Chess Society and was pleasantly surprised to meet the vibrant Pokémon GO community in Edinburgh.

I owe a particular debt of gratitude to Tomasz and his family for their swift help in taking me to the Royal Infirmary when I was in a vulnerable medical situation. I thank the doctors and staff at the Royal Infirmary, St. John's, and Lauriston for their care. However, this difficult time brought a wonderful connection into my life, leading me to meet Sai Akka, Krishna Bava, Charvik, and Bhaavya, who became my family away from home. Thanks also to Mary Akka for connecting me to Sai Akka.

I am incredibly glad to have met Nijesh and Nickil during my final six months, as they helped me make some of my best memories. I will always cherish our discussions, which sometimes went well past midnight—often ignoring Nijesh's protests that he needed to sleep, even though he was usually the one keeping the conversation alive.

This section would not be complete without mentioning Vishu, Aditya, Rohith, Satish, Uday, KVN, Chetan, Jagadeesh, Raghuveer, and many more who have been a constant in my life for the past 10+ years.

Finally, I would like to thank my parents, Aravinda and Subba Rao, and my sister, Pranathi, for their unwavering support and for the sacrifices they made to ensure I had the life and opportunities that led me here.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Mohan Sai Teja Dantam*)

Contents

Chapter 1

Introduction

1.1 Motivation

Sequential decision making concerns with the setting where an agent or a program is to repeatedly resolve non-deterministic choices to achieve a goal, possibly in an optimal fashion if one exists. For example, consider a vacuum cleaner robot trying to locate and pick up trash automatically in a room. We are interested in the planning decisions that the robot makes and assume it is perfectly capable of discerning rubbish from obstacles which are present on the floor. The path that the robot takes ideally would be such that (i) it avoids all obstacles on the floor and (ii) it travels the minimum distance to pick up all the trash so as to optimize for the power consumption. Violating (i) or (ii) but already in this toy example, one can see that there is more than one property we are interested in that we have to simultaneously satisfy. Let's take a simpler example of reaching a target state in a weighted directed graph with minimum cost. This can be thought of as an abstraction for taking the shortest path to go home from work. Once again, notice that we really have two properties that we are interested in, (i) reaching the target state and (ii) minimizing the cost.

In both of the above examples, once a decision control is fixed, we assume certainty about what happens next. This is clearly not always the case because the agent is part of a bigger *environment* whose state one cannot always predict. This leads us to include stochasticity into the model. Not only that, but in the second example it could be that there is only a single narrow road and there is another person whose house is in the same direction as yours, making this a competitive process.

In this thesis, we consider problems where the decisions will have to be taken in the presence of an environment (random agent) and possibly an additional adversarial agent. Furthermore, we deal with the case where the choice of non-determinism is what edge to take from a vertex, *i.e.*, the arena is a finite graph where some vertices are controlled by the player. We have more than one property that we are interested in satisfying.

More precisely, the models we look at are finite-state *Markov Decision Processes* ([Put94]) and *2-player turn based stochastic games with perfect information* ([Con92]). We describe them in more detail. Before that, we look at some of the broader applications of Algorithmic Game Theory.

1.2 Scientific Context

The mathematical frameworks and algorithms developed in this thesis are situated within a broader paradigm shift in Computer Science: the transition from analyzing static, isolated computations to modeling dynamic interactions between autonomous agents. While the technical contributions of this work focus on the rigorous complexity analysis of stochastic games, the importance of these results is best understood through the diverse applications of game theory in modern computational systems.

The Success of Algorithmic Game Theory

Historically, Computer Science concerned itself with the efficient execution of instructions by a single machine. However, the rise of the internet necessitated a framework for analyzing systems where "correctness" is determined by the equilibrium states of multiple self-interested entities [NRTV07]. **Algorithmic Game Theory (AGT)** has since become a fundamental pillar for optimization, with two prominent success stories demonstrating its real-world impact:

- **Mechanism Design and Auctions:** Perhaps the most commercially significant application of AGT is the design of auctions. This ranges from the *Generalized Second-Price (GSP)* auctions that power online advertising (e.g., Google AdWords), generating billions in revenue by incentivizing truthful bidding [EOS07, Var07], to the combinatorial auctions used by governments for allocating radio spectrum [Mil00].

- **Matching Markets:** Beyond monetary markets, AGT has revolutionized markets where money is repugnant or illegal. The theory of stable matching has been successfully deployed to redesign the *National Resident Matching Program (NRMP)* for doctors [Rot84] and to organize *Kidney Exchanges*, where cycles of incompatible patient-donor pairs are matched to save lives [RSÜ04].

Stochastic Games and Formal Verification

While the applications above typically focus on equilibrium (predicting how rational agents *will* behave), the field of **Formal Verification** uses game theory to establish correctness (guaranteeing how a system *must* behave). Here, the "game" is played between a system (the controller) and its environment (the adversary). When these environments exhibit uncertainty—due to sensor noise, random faults, or market fluctuations—**Stochastic Games** and **Markov Decision Processes (MDPs)** become the necessary modeling tools.

A central challenge in modern systems engineering is that real-world problems are rarely mono-dimensional. An autonomous agent does not simply wish to reach a target; it must do so while minimizing fuel consumption, avoiding unsafe regions, and maximizing scientific return. This necessitates the study of **Games and MDPs with Multiple Objectives**. Unlike single-objective optimization, multi-objective settings require analyzing trade-offs to find *Pareto-optimal* strategies or satisfying a conjunction of constraints (e.g., "maximize reward R subject to energy constraint E") [EKVY08, CMH06].

Tools and Applications in Reliable AI

The practical application of these theoretical concepts is driven by mature software tools. **Probabilistic Model Checkers** such as **PRISM** [KNP11] and **Storm** [HJK+22] rely on the algorithmic foundations of stochastic games to automatically verify quantitative properties of complex systems.

These tools are increasingly vital in the intersection of Formal Methods and **Machine Learning**. For example, techniques like **Shielding** use winning strategies from stochastic safety games to monitor Reinforcement Learning (RL) agents, overriding actions that would violate safety constraints [ABE+18]. By providing the theoretical bounds and algorithms for these multi-objective games, this thesis

supports the development of AI systems that are not only intelligent but provably safe.

1.3 Model

We consider 2-player simple stochastic games on graphs. One way to look at this graph is as an abstraction of some transition system. The vertex set is partitioned into 3 sets, one for each player $\Box$ and $\Diamond$ and remaining vertices belong to the environment $\bigcirc$. The game starts at some vertex and continues for infinite duration. When the play is at some vertex, the player to whom that vertex belongs chooses the edge. If the vertex is a chance node, then an edge is chosen with predefined distribution. After infinite time, both players get a payoff based on the infinite run and the valuation of a given objective on this infinite run. We assume that the game is zero-sum, *i.e.*, any positive payoff for one player is a negative for other. Since the graph is inherently stochastic, players want to maximize their expected payoff. Because the game is zero-sum, maximizing your own payoff and minimizing the adversary's payoff are equivalent notions. So, we can consider a single objective function which the players want to either maximize or minimize. We call the players Max and Min using this point of view.

The above model is quite general and subsumes both Markov Decision Processes (MDPs) and deterministic 2-player zero-sum games on graphs. Deterministic 2-player games commonly occur in the synthesis for reactive systems [PR89, RW87]. Markov Decision Processes, which have applications in various fields [SB18, Sch02], can be seen as games where one of the players has no vertices with a non-trivial choice.

A further generalization of simple stochastic games is the case of concurrent games [Sha53] where the vertex set is not partitioned and the next state depends on the choice of both players who choose an action at every step. Although termed here as a generalization, it would be historically more accurate to phrase simple stochastic games as a simpler case of concurrent games. However, we do not work with concurrent games, so stochastic games in this thesis always refer to turn-based stochastic games.

1.4 Objectives

We look at some commonly studied objectives for Games on Graphs. Most of the objectives we consider can simply be defined as a subset of the set of all possible infinite plays. The payoff function is 1 if it belongs to the subset, 0 otherwise.

Parity. Parity is a characterisitic objective. The states of the game are assigned colors and an infinite run is accepted into the set iff the maximum color occuring infinitely often in the run is even. Parity can also defined in terms of minimum color and requiring that the color is odd instead of even. All definitions are equivalent as one can simply change the coloring function to go from one definition to another. It is expressive enough to express all ω-regular objectives. In fact, one of the standard approaches for synthesis for LTL specifications is to construct an equivalent parity automaton and check for emptiness [BCJ18, Section 3.7].

All the following objectives will be defined w.r.t some weight function on the edges.

Discounted MeanPayoff. Given a discount factor $0 \leq \gamma < 1$, the payoff of an infinite run is the γ-discounted sum of weights on edges taken during the run.

A characterisitic version of this objective could be defined by considering a threshold k and accepting a run iff the payoff is $>$ or $\geq k$.

MeanPayoff. First considered by Gillette [Gil57], the objective is to maximize the expected long term average reward. Since, the long term average is a limit which may not exist, one usually considers either the limit infimum or limit supremum as the payoff.

Similar to discounted meanpayoff, one could also define a characterisitic variant of meanpayoff based on some threshold k.

Energy/ Termination. Introduced in [CDAHS03], energy objectives arise out of the need for checking compatibility between different components of a system each of which consumes and/or generates a certain resource. Compatibility w.r.t the given resource would then imply that the system is self-sufficient in this resource.

In games with energy objective, the game starts with a given amount of initial energy and transitions either modify this value by adding the weight on the edge. The player tries to keep the energy level non-negative.

Viewed dually from the point of view of adversary, these can also be seen as termination games [BBE10a], where the goal is to get rid of the resource completely.

1.5 Strategy & Computational Complexity

We saw earlier that given a characterisitic objective, Max wants to maximize the probability that the play belongs to the objective, while Min wants to minimize this probability. This then leads to the question "*Given an objective, what is the best way to play?*". A 'way to play' or strategy for now can be understood roughly as choosing an edge at every step where it is the player's choice. On first glance, it is not even clear that there has to be a best strategy as the payoff is dependent on the entire infinite run. In fact, it turns out that for termination, there is no best strategy [BBE$^+$10c] even in the case of MDPs. So, alternatively, one can ask "*Given δ, is there a strategy which guarantees Max a payoff of at least δ against all strategies of Min?*" Deciding these questions and understanding the time and the space taken by the best algorithm is fundamental to understand the hardness of the objectives and models.

All the objectives we consider are sufficiently nice so that they are Borel measurable and hence *weakly-determined* [MS03, Mar98]. This means, although there might not be a 'best' strategy from a state s, there is a number ν_s such that no strategy of Max can guarantee a payoff $> \nu_s + \varepsilon$ and no strategy of Min can force a payoff $< \nu_s - \varepsilon$ for every $\varepsilon > 0$. This can be thought of as follows: *While there may not be best strategies, both players can play something which is arbitrarily close to it.* These will be termed ε-optimal strategies and finding an algorithm which takes the minimum time to compute ν_s and these strategies is also an important question considered in the literature.

From the definition of value, it is clear if there is some strategy for Max, which achieves a payoff of at least ν_s against every strategy of Min, then it is a 'best' strategy for Max. Such a strategy will be termed as an optimal strategy. Deciding the *existence of optimal strategies* is another question one could consider in this context.

If both players have best strategies, then it is easy to see that payoff for Max would be equal to the value of the start state. When the value of a state is 1, and is achieved by some Max strategy σ, σ is called an almost surely winning strategy for Max. Dually, when the value of a state is 0, and this is achieved by some Min strategy π, it is an almost surely winning strategy for Min. These are important to consider, since such a strategy guarantees the satisfaction of some properties, irrespective of the randomness and the strategy of the other player. So one can then ask *Given a state, is there an almost surely winning strategy for Max or for Min?*

For all the questions above which also involve computing a strategy, one could also ask questions about the 'complexity' of the computed strategy *i.e.*, the memory required and whether the strategy has to randomized or the tradeoff between memory and randomization. Finite memory strategies are usually defined as automata which take the states as alphabet and output the edge to take if it is the players' turn. The rough definition of strategies we gave above, while good as a starting point, doesn't encompass all possible strategies. In fact, the weak determinacy of Borel measurable objectives is only true if one considers randomized strategies. Roughly speaking, at every turn of the player, the definition is generalized so that the player can give a distribution on edges from the state instead of choosing a single edge. The class of randomized strategies can express behaviors which are otherwise not possible by deterministic strategies [MR22].

1.6 Contributions

We briefly summarize the main contributions of this thesis. More explicit description is provided at the beginning of the respective chapters.

Chapter 3: Approximating Energy-Parity. We present an algorithm to approximate the value of a given configuration in games (Theorem 3.1). Moreover, ε-optimal strategies for either player require at most doubly exponential number of memory modes in the case of unary rewards.

Chapter 4: Finiteness of Energy-MeanPayoff strategies. We show that finite memory suffices for almost surely winning strategies for the Energy-MeanPayoff objective in Markov Decision Processes (Theorem 4.1). We show that exponential

memory is sufficient (even for deterministic strategies) and necessary (even for randomized strategies).

Chapter 5: Lifting for randomized strategies. We show a tight lower bound of exponential memory blowup when lifting finite-memory strategies from MDPs to stochastic games for shift-invariant inverse-submixing objectives (Theorem 5.2). We show that optimal strategies for Meanpayoff-Parity only require a linear number of memory modes, even though it is shift-invariant and inverse-submixing (Theorem 5.5). A different lifting with orthogonal assumptions does not generalize to randomized strategies (Proposition 5.23).

1.7 Outline of the Thesis

Chapter 2 introduces some basic notation, definitions and results which are common and used in all chapters. It also contains a section on some results in Markov chains with rewards which is used in the proof of Lemma 4.11, but the reason we include it here instead of in Chapter 4 is twofold. One, it is too long and distracts from the main argument for the results there. Two, it could be of independent interest.

The following three chapters start with an overview which briefly explains the contributions in the chapter, followed by an introduction which contains any additional preliminaries needed specific for the chapter, related work, and our contributions. **All the results are the product of close collaboration with my advisor from the formation of initial idea to writing up the results.**

Chapter 3 contains results about the computation and complexity for approximating the value of a state for the Energy-Parity objective in stochastic games. Along with this, we also compute ε-optimal strategies for both players.

Chapter 4 looks at the Energy-MeanPayoff objective in MDPs, specifically the strategy complexity for the almost surely winning strategies, and the computational complexity of finding the set of almost surely winning states.

Chapter 5 looks at hypotheses which allow one to lift strategies from MDPs to stochastic games and analyses the extension to randomized strategies. Additionally, it also solves the strategy complexity for MeanPayoff-Parity in stochastic games.

Chapter 6 summarizes the results and looks at possible future research directions and questions to explore.

Chapter 2

Notation

2.1 Preliminaries

Let $\mathbb{Z}$ (resp. $\mathbb{Z}_+, \mathbb{N}, \mathbb{Q}$) denote the set of integers (resp. positive, non-negative integers, rationals). The 'size' of an integer n or a rational $q = \frac{m}{n}(\gcd(m,n) = 1)$ is defined in a natural way assuming binary representation. $\|n\| \stackrel{\text{def}}{=} \lceil \log_2(|n|+1) \rceil + 1$ and $\|q\| \stackrel{\text{def}}{=} \|m\| + \|n\| + 1$. A *probability distribution* over a countable set S is a function $f : S \to [0,1]$ with $\sum_{s \in S} f(s) = 1$. $\mathsf{supp}(f) \stackrel{\text{def}}{=} \{s \mid f(s) > 0\}$ denotes the support of f and $\mathcal{D}(S)$ is the set of all probability distributions over S. If S is finite and $f(S) \subseteq \mathbb{Q}$, we define *bit size* of f to be $\text{bits}(f) \stackrel{\text{def}}{=} \sum_{s \in S}(\|f(s)\|)$. One can extend this to probabilistic functions with rational probabilities of the form $p : S \to \mathcal{D}(T)$ when both S and T are finite by defining $\text{bits}(p) \stackrel{\text{def}}{=} \sum_{s \in S} \text{bits}(p(s))$.

Games, MDPs and Markov chains. A *Simple Stochastic Game* (SSG) is a finite-state 2-player turn-based perfect-information stochastic game $\mathcal{G} = (S, (S_\Box, S_\Diamond, S_\bigcirc), E, P)$ where the finite set of states S is partitioned into $S_\Box$(*Max* states), $S_\Diamond$(*Min* states), and $S_\bigcirc$(chance states). If $S_\bigcirc = \emptyset$ then it is called a deterministic 2-player game. Let $E \subseteq S \times S$ be the transition relation. We write $s \longrightarrow s'$ if $(s, s') \in E$ and assume that $\mathsf{Succ}(s) \stackrel{\text{def}}{=} \{s' \mid sEs'\} \neq \emptyset$ for every state s. The *probability function* P assigns each random state $s \in S_\bigcirc$ a distribution over its successor states, *i.e.*, $P(s) \in \mathcal{D}(\mathsf{Succ}(s))$. For ease of presentation, we extend the domain of P to $S^*S_\bigcirc$ by $P(\rho s) \stackrel{\text{def}}{=} P(s)$ for all $\rho s \in S^+S_\bigcirc$. An *MDP* is a game where one of the two players does not control any states. An MDP is *maximizing* (resp. *minimizing*) iff $S_\Diamond = \emptyset$ (resp. $S_\Box = \emptyset$). A *Markov chain* is a game with only random states, *i.e.*, $S_\Box = S_\Diamond = \emptyset$.

Strategies. A *play* is an infinite sequence $s_0 s_1 \ldots \in S^\omega$ such that $s_i \longrightarrow s_{i+1}$ for all $i \geq 0$. A *path* is a finite prefix of a play. Let $\mathit{Plays}(\mathcal{G}) \stackrel{\text{def}}{=} \left\{ (s_i)_{i \in \mathbb{N}} \mid s_i \longrightarrow s_{i+1} \right\}$ denote the set of all possible plays. A strategy of the player $\Box$ ($\Diamond$) is a function $\sigma : S^* S_\Box \to \mathcal{D}(S)$ ($\pi : S^* S_\Diamond \to \mathcal{D}(S)$) that assigns to every path $ws \in S^* S_\Box$ ($\in S^* S_\Diamond$) a probability distribution over the successors of s. If these distributions are always Dirac then the strategy is called *deterministic*, otherwise it is called *randomized*. Strategies defined in this fashion are called *behavioral* strategies. The set of all strategies of player $\Box$ and $\Diamond$ in $\mathcal{G}$ is denoted by $\Sigma_\mathcal{G}$ and $\Pi_\mathcal{G}$, respectively. A play/path $s_0 s_1 \ldots$ is compatible with a pair of strategies (σ, π) if $s_{i+1} \in \mathsf{supp}(\sigma(s_0 \ldots s_i))$ whenever $s_i \in S_\Box$ and $s_{i+1} \in \mathsf{supp}(\pi(s_0 \ldots s_i))$ whenever $s_i \in S_\Diamond$.

An alternative way to define strategies is via *Mealy Machines* [Mea55]. In this notion, a strategy for a player $\odot \in \{\Box, \Diamond\}$ is a tuple $\tau = (\mathsf{M}, \mathsf{m}_0, \mathsf{upd}, \mathsf{nxt})$ where M is the set of memory modes, m_0 is the initial memory mode, the next/ successor function $\mathsf{nxt} \colon \mathsf{M} \times S_\odot \to \mathcal{D}(S)$ chooses a (distribution over) successor states based on the current memory mode and state and the update function $\mathsf{upd} \colon \mathsf{M} \times E \to \mathcal{D}(\mathsf{M})$ updates the memory mode upon observing a transition.

To convert a strategy $\tau \colon S^* S_\odot \to \mathcal{D}(S)$ given in functional form to a Mealy machine, one can trivially consider M to be the set of all possible finite words S^*. The initial memory mode would then simply be the empty prefix ε and nxt will be defined as $\mathsf{nxt} \colon (w, s_\odot) \mapsto \tau(w s_\odot)$. The update function will be deterministic, $\mathsf{upd} \colon (w, (s, s')) \mapsto ws'$. The mapping is only defined if the last state of w is s as otherwise it is not possible to take an edge. For the other direction, one can refer to [MR22]. As can be observed from the former conversion, *when considering all strategies, deterministic updates suffice.*

We define finite-memory strategies w.r.t. Mealy machine definition. Finite-memory strategies are a subclass of strategies which use a finite set M of memory modes. Unlike the case for general strategies, *randomized updates are strictly more expressive than deterministic updates for finite- memory strategies* [MR22]. The set of all finite-memory Max (resp. Min) strategies in $\mathcal{G}$ is denoted by $\Sigma_f^\mathcal{G}$(resp. $\Pi_f^\mathcal{G}$). Let $\tau[\mathsf{m}]$ denote the finite-memory strategy τ starting in memory mode m instead of m_0.

Strategies with memory $|\mathsf{M}| = 1$ are called *memoryless*. Finite-memory strategies with deterministic/ randomized next function and deterministic/ randomized update function are called FDD, FRD, FDR, FRR, while memoryless

deterministic (resp. randomized) strategies are called MD (resp. MR). In finite memory strategies, the first D/R refers to randomization in the next function and second D/R refers to randomization in the update function. This notation is inspired from [MR22, Figure 1.1] and adapted to our setting since we do not consider randomization over the initial memory mode.

Note that the definition by Mealy machines is not the only way to model strategies and other possibilities such as decision trees [BCKT18], strategy machines [Gel14] etc. are also studied in the literature. Even in the Mealy machine model, there are different classes which differ based on the information they take as input [Kop08].

We are generally interested in the strategies of Max so whenever we talk about a strategy and do not explicitly mention the player, it is assumed to be Max.

size of a strategy: When a general finite-memory strategy τ is given by $(\mathsf{M}, \mathsf{m}_0, \mathsf{upd}, \mathsf{nxt})$, The size of τ is defined as $\|\tau\| \stackrel{\text{def}}{=} |\mathsf{M}|$. We are also interested in the total bit size of the probabilities used in the strategies which use rational probabilities. The total bit size in such case is defined as $\text{bits}(\tau) \stackrel{\text{def}}{=} \text{bits}(\mathsf{nxt}) + \text{bits}(\mathsf{upd})$.

Measure. A game $\mathcal{G}$ with initial state s_0 and strategies (σ, π) yields a probability space $(s_0 S^\omega, \mathcal{F}_{s_0}, \mathcal{P}^{\mathcal{G}}_{\sigma,\pi,s_0})$ where $\mathcal{F}_{s_0}$ is the σ-algebra generated by the cylinder sets $s_0 s_1 \dots s_n S^\omega$ for $n \geq 0$. The probability measure $\mathcal{P}^{\mathcal{G}}_{\sigma,\pi,s_0}$ is first defined on the cylinder sets. For $\rho = s_0 \dots s_n$, let $\mathcal{P}^{\mathcal{G}}_{\sigma,\pi,s_0}(\rho) \stackrel{\text{def}}{=} 0$ if ρ is not compatible with σ, π and otherwise $\mathcal{P}^{\mathcal{G}}_{\sigma,\pi,s_0}(\rho S^\omega) \stackrel{\text{def}}{=} \prod_{i=0}^{n-1} \tau(s_0 \dots s_i)(s_{i+1})$ where τ is σ or π or P depending on whether $s_i \in S_\Box$ or $S_\Diamond$ or $S_\bigcirc$, respectively. By Carathéodory's extension theorem [Bil08], this defines a unique probability measure on the σ-algebra.

Objectives and Payoff functions. General objectives are defined by real-valued measurable functions $v \colon s_0 S^\omega \to \mathbb{R}$, and we write $\mathcal{E}(\cdot)$ for the expectation w.r.t. $\mathcal{P}$ and v. For event-based objectives, v is just the indicator function of a measurable set of plays $\mathtt{O} \subseteq S^\omega$, *i.e.*, we consider the probability that plays belong to $\mathtt{O}$.

An objective v is called *shift-invariant* iff for all finite paths ρ and infinite plays $\rho' \in S^\omega$, we have $v(\rho\rho') = v(\rho')$. It is called *submixing* iff for all sequences of finite non-empty words u_0, w_0, u_1, $w_1 \dots$ we have $v(u_0 w_0 u_1 w_1 \dots) \leq \max(v(u_0 u_1 \dots), v(w_0 w_1 \dots))$, and symmetrically, it is *inverse-submixing* iff we

have $v(u_0w_0u_1w_1\ldots) \geq \min(v(u_0u_1\ldots), v(w_0w_1\ldots))$ [GK23].

Reachability. We use the syntax and semantics of the LTL operators [CGP99] F (eventually) and G (always) to specify some conditions on plays. A *reachability objective* is defined by a set of target states $T \subseteq S$. A play $\rho = s_0s_1\ldots$ belongs to $\mathsf{F}\,T$ iff $\exists i \in \mathbb{N}\, s_i \in T$. Similarly, ρ belongs to $\mathsf{F}^{\leq n}T$ (resp. $\mathsf{F}^{\geq n}T$) iff $\exists i \leq n$ (resp. $i \geq n$) such that $s_i \in T$. Dually, the *safety* objective $\mathsf{G}\,T$ consists of all plays which never leave T. We have $\mathsf{G}\,T = \neg\mathsf{F}\neg T$.

Parity. A parity objective is defined via a bounded function $Col : S \to \mathbb{N}$ that assigns non-negative priorities (aka colors) to states. Given an infinite play $\rho = s_0s_1\ldots$, let $\mathrm{Inf}(\rho)$ denote the set of numbers that occur infinitely often in the sequence $Col(s_0)Col(s_1)\ldots$. A play ρ satisfies *even parity* w.r.t. Col iff the maximum [1] of $\mathrm{Inf}(\rho)$ is even. Otherwise, ρ satisfies *odd parity*. The objective even parity is denoted by $\mathtt{EPAR}(Col)$ and odd parity is denoted by $\mathtt{OPAR}(Col)$. Most of the time, we implicitly assume that the coloring function is known and just write $\mathtt{EPAR}$ and $\mathtt{OPAR}$. Observe that, given any coloring Col, we have $\overline{\mathtt{EPAR}} = \mathtt{OPAR}$ and $\mathtt{OPAR}(Col) = \mathtt{EPAR}(Col + 1)$ where $Col + 1$ is the function which adds 1 to the color of every state. This justifies to consider only one of the even/odd parity objectives, but, for the sake of clarity, we distinguish these objectives wherever necessary. For any subset T and color c, we denote by $T(c)$, states in T with color c.

Reward based objectives. Let $r : E \to \{-R, \ldots, 0, \ldots, R\}$ be a bounded function that assigns rewards to transitions. If $s \longrightarrow s'$ and $r((s, s')) = c$, we write $s \xrightarrow{c} s'$. Let $\rho = s_0 \xrightarrow{c_0} s_1 \xrightarrow{c_1} \ldots$ be a play. Since edges in our case are represented as pairs of states, we often write $r(s, s')$ instead of $r((s, s'))$ for ease of notation. We say that ρ satisfies

1. The k-*energy* objective $\mathtt{EN}(k)$ iff $\left(k + \sum_{i=0}^{n-1} c_i\right) > 0$ for all $n \geq 0$.

2. The l-*storage condition* if $l + \sum_{i=m}^{n-1} c_i \geq 0$ holds for every infix $s_m \xrightarrow{c_m} s_{m+1}\ldots s_n$ of the play. Let $\mathtt{ST}k, l)$ denote the set of plays that satisfy both the k-energy and the l-storage condition. Let $\mathtt{ST}k) \stackrel{\text{def}}{=} \bigcup_l \mathtt{ST}k, l)$. Clearly, $\mathtt{ST}k) \subseteq \mathtt{EN}(k)$.

3. k-*Termination* $\mathtt{Term}(k)$ iff there exists $n \geq 0$ such that $\left(k + \sum_{i=0}^{n-1} c_i\right) \leq 0$.

[1] exists since Col is bounded

4. *Limit* objective $\texttt{LimInf}(\triangleright z)$ iff $\left(\liminf_{n\to\infty} \sum_{i=0}^{n-1} c_i\right) \triangleright\ z$ for $\triangleright \in \{<, \leq, =, \geq, >\}$ and $z \in \mathbb{R} \cup \{\infty, -\infty\}$ and similarly for $\texttt{LimSup}(\triangleright z)$.

5. *Mean payoff* $\texttt{MP}\triangleright c$ for some constant $c \in \mathbb{R}$ iff $\left(\liminf_{n\to\infty} \frac{1}{n}\sum_{i=0}^{n-1} c_i\right) \triangleright\ c$.

Multidimensional reward-based objectives. For a d-dimensional real vector $\boldsymbol{\mu}$, let μ_i denote the i^{th} component of $\boldsymbol{\mu}$ for $1 \leq i \leq d$. Given two vectors $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{R}^d$, $\sim \in \{<, \leq, >, \geq, =\}$ we say $\boldsymbol{\mu} \sim \boldsymbol{\nu}$ if $\mu_i \sim \nu_i$ for every i. In particular, $\boldsymbol{\mu} > \mathbf{0}$ means that *every* component of $\boldsymbol{\mu}$ is strictly greater than 0. For a multidimensional reward function $\boldsymbol{r} : E \to [-R, R]^d$, we can consider any boolean combination of reward based objectives using any components of $\boldsymbol{r}$. For instance, $\texttt{O}_1 = \texttt{EN}_1(k) \cap \texttt{MP}_2(> 0)$ denotes the objective that contains all runs that satisfy $\texttt{EN}(k)$ in the 1^{st} dimension and $\texttt{MP}(> 0)$ in the 2^{nd} one. We denote conjunctions of the same objective across different dimensions in vectorized form, with the dimension information in the subscript. Therefore, $\texttt{EN}_{[a,b]}(\boldsymbol{k}) \cap \texttt{MP}_{[c,d]}(\boldsymbol{> x})$ denotes the runs where the $\texttt{EN}_i(k_i)$ objective is satisfied for each $i \in [a, b]$ and the $\texttt{MP}_j(> x_j)$ objective is satisfied for each $j \in [c, d]$.

Sometimes, we also consider the complement of an objective $\texttt{O}$ which will be denoted by $\overline{\texttt{O}}$.

size of a game: The size of a game $\mathcal{G}$ with an objective $\texttt{O}$ can be thought of as the number of bits required to describe the states, transitions and probabilities used by $\mathcal{G}$ along with the description of $\texttt{O}$. Often, the objective will be implicit from the context and we simply write $\|\mathcal{G}\|$.

Determinacy. Given an objective v and a game $\mathcal{G}$, state s has value (w.r.t v) iff

$$\sup_{\sigma\in\Sigma_\mathcal{G}} \inf_{\pi\in\Pi_\mathcal{G}} \mathcal{E}^\mathcal{G}_{\sigma,\pi,s}(v) = \inf_{\pi\in\Pi_\mathcal{G}} \sup_{\sigma\in\Sigma_\mathcal{G}} \mathcal{E}^\mathcal{G}_{\sigma,\pi,s}(v).$$

If s has value then $\texttt{val}^\mathcal{G}_v(s)$ denotes the value of s defined by the above equality. A game with an objective is called *weakly determined* if every state has value. Stochastic games with Borel objectives are weakly determined [MS03, Mar98]. Our objectives above are Borel, hence any boolean combination of them is also weakly determined. For $\varepsilon > 0$ and state s, a strategy

1. $\sigma \in \Sigma_\mathcal{G}$ is ε-optimal (maximizing) iff $\mathcal{E}^\mathcal{G}_{\sigma,\pi,s}(v) \geq \texttt{val}^\mathcal{G}_v(s) - \varepsilon$ for all $\pi \in \Pi_\mathcal{G}$.

2. $\pi \in \Pi_\mathcal{G}$ is ε-optimal (minimizing) iff $\mathcal{E}^\mathcal{G}_{\sigma,\pi,s}(v) \leq \texttt{val}^\mathcal{G}_v(s) + \varepsilon$ for all $\sigma \in \Sigma_\mathcal{G}$.

A 0-optimal strategy is called *optimal*. An MD strategy is called *uniformly* ε-optimal (resp. uniformly optimal) if it is so from every start state. Given an event-based objective $\mathtt{O}$, an optimal strategy for player $\Box$ from state s is *almost surely* winning if $\mathtt{val}_{\mathtt{O}}^{\mathcal{G}}(s) = 1$.

By $\mathtt{AS}_{\Box}^{\mathcal{G}}(\mathtt{O})$ we denote the set of states that have an almost surely winning strategy for $\Box$ for objective $\mathtt{O}$. For ease of presentation, we drop subscripts and superscripts wherever possible if they are clear from the context.

Attractors, Traps and Subgames. A non-empty set $H \subseteq S$ defines a *subgame* iff for all $s \in H \cap S_{\bigcirc}$, $\mathsf{Succ}(s) \subseteq H$ and every state in H has at least one successor in H. The resulting subgame which is well-defined is denoted by $\mathcal{G}[H]$. For a set $T \subseteq S$, the *positive attractor* for player $\odot$ in game $\mathcal{G}$, denoted by $\mathtt{Attr}_{\odot}(T, \mathcal{G})$ is the set of states s where $\odot$ has a strategy to ensure that $\mathsf{F}T$ is satisfied with positive probability when starting from s. For any given T, the positive attractor set and a uniform MD strategy $\tau_{\mathtt{Attr}}$ which satisfies this can be computed in polynomial time. A set $U \subseteq S$ is a *trap* for Min iff for every Min/ random state in U, the successors are always in U and there is at least one successor in U for every Max state. One can similarly define a trap for Max. Note that by definition a trap for either player is a subgame. Also, for any set T, $S \setminus \mathtt{Attr}_{\odot}(T, \mathcal{G})$ is a trap for $\odot$ if it is non-empty.

Remark 2.1. For finite-state SSGs and the following objectives there exist optimal MD strategies for both players. Moreover, if the SSG is just a maximizing MDP then the set of states that are almost surely winning for Max can be computed in polynomial time.

1. $\mathsf{F}\,T$ [Con92]
2. $\mathtt{LimInf}(\rhd \pm \infty)$, $\mathtt{LimSup}(\rhd \pm \infty)$, $\mathtt{MP}\,\rhd 0$ [BBE10a, Prop. 1]
3. $\mathtt{EPAR}$ [Zie98]

Induced MDP. We sometimes consider situations where the strategy of one of the players is fixed in advance. This results in a MDP of the other player. If the strategy that is fixed also has memory, this results in an MDP whose states also have the memory of the fixed strategy as part of the state. Formally,

▶ **Definition 2.2** (Induced maximizing MDP)**.** *Given a simple stochastic game* $\mathcal{G} = (S, (S_\Box, S_\Diamond, S_\bigcirc), E, P)$ *and a finite-memory (FR) strategy* $\pi = (\mathsf{M}, \mathsf{m}_0, \mathsf{upd}, \mathsf{nxt})$ *for Min let* $\mathcal{G}_\pi$ *be the maximizing MDP with state space* $\mathsf{M} \times (S \uplus E)$ *obtained by fixing Min's choices according to* π*. The transition rules* $\longrightarrow'$ *and the partition of the states in the derived MDP* $\mathcal{G}_\pi$ *are given as follows.*

1. *Every state in* $\mathsf{M} \times E$ *is a randomized state with*

$$P'((\mathsf{m}, (s, s'))) = \mathsf{upd}((\mathsf{m}, (s, s'))) \cdot \delta_{(\cdot, s')}$$

   i.e.*, Min determines the next memory mode.*

2. *If* $s \in S_\Box$*, every* (m, s) *is a Max state and for every* $s \longrightarrow s'$*,* $\mathsf{m} \in \mathsf{M}$*, we have* $(\mathsf{m}, s) \longrightarrow' (\mathsf{m}, (s, s'))$*,* i.e.*, Max determines the successor state.*

3. *Similarly if* $s \in S_\bigcirc$*, every* (m, s) *is a random state and for every* $s \longrightarrow s'$*,* $\mathsf{m} \in \mathsf{M}$ *we have* $(\mathsf{m}, s) \longrightarrow' (\mathsf{m}, (s, s'))$ *and* $P'((\mathsf{m}, s))((\mathsf{m}, s')) = P(s)(s')$*,* i.e.*, transition probabilities are inherited.*

4. *If* $s \in S_\Diamond$*, every* (m, s) *is a random state and for every* $s' \in \mathsf{supp}(\mathsf{nxt}((\mathsf{m}, s)))$ *we have* $(\mathsf{m}, s) \longrightarrow' (\mathsf{m}, (s, s'))$ *and* $P'((\mathsf{m}, s))((\mathsf{m}, (s, s'))) = \mathsf{nxt}((\mathsf{m}, s))(s')$*,* i.e.*, Min chooses the successor state according to* nxt *function of* π*.*

In the dual case where a FR strategy σ *for Max is fixed, we obtain a minimizing MDP* $\mathcal{G}^\sigma$*. The construction is the same as above, with the roles of Min and Max swapped.*

Given an infinite run $\rho = s_0 \xrightarrow{c_0} s_1 \xrightarrow{c_1} \ldots$, let $X_n(\rho) \stackrel{\text{def}}{=} s_n$ denote the n-th state. Let Y_n be the sum of the rewards in the first n steps, *i.e.*, $Y_n(\rho) \stackrel{\text{def}}{=} \sum_{i=0}^{n-1} c_i$. These become random variables once an initial distribution and a strategy are fixed.

2.2 Bounds in Markov Chains

This section shows some generic results for Markov chains with transition rewards. We show bounds on the expected arrival time (aka first passage time) of situations when the total reward reaches particular levels, under the condition that the total reward is truncated to remain inside some interval $[a, b]$. E.g., the total reward might hit the upper limit b many times (and be truncated there) before arriving

at the lower level a for the first time. These bounds are later applied to Markov chains obtained by fixing certain finite-memory strategies in MDPs, and they are used in the proof of Lemma 4.11.

To establish these bounds formally, we rely on techniques from martingale theory. For completeness and to fix our notation, we briefly recall the standard definitions regarding filtrations, stopping times, and martingales in the context of a probability space $(\Omega, \mathcal{F}, \mathcal{P})$.

▶ **Definition 2.3** (Filtration). *A* filtration *is a sequence of sub-σ-algebras $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}$. Intuitively, $\mathcal{F}_n$ represents the accumulated information available at time n.*

▶ **Definition 2.4** (Stopping Time). *A random variable $T : \Omega \to \mathbb{N} \cup \{\infty\}$ is a* stopping time *with respect to a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ if for every $n \in \mathbb{N}$, the event $\{T \leq n\}$ is in $\mathcal{F}_n$. Equivalently, the decision to stop at time n is determined entirely by the information available at time n (*i.e., *$\{T = n\} \in \mathcal{F}_n$).*

▶ **Definition 2.5** (Martingale). *A sequence of random variables $M_0, M_1, \ldots$ is a* martingale *with respect to a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ if, for all $n \geq 0$, M_n is $\mathcal{F}_n$-measurable, $\mathcal{E}(|M_n|) < \infty$, and*

$$\mathcal{E}(M_{n+1} \mid \mathcal{F}_n) = M_n.$$

▶ **Theorem 2.6** (Optional Stopping Theorem). *Let M be a martingale and T be a stopping time with respect to a filtration $\mathcal{F}_n$. If $\mathcal{P}(T < \infty) = 1$ and one of the following conditions holds:*

1. *T is bounded (i.e., there exists K such that $T \leq K$ almost surely); or*
2. *$\mathcal{E}(T) < \infty$ and there exists a constant c such that $|M_{n+1} - M_n| \leq c$ almost surely (Bounded Increments); or*
3. *M is bounded (i.e., there exists c such that $|M_n| \leq c$ almost surely);*

then $\mathcal{E}(M_T) = \mathcal{E}(M_0)$. Standard references include [Bil08].

In the specific setting of a Markov chain $\mathcal{A}$ with states $X_0, X_1, \ldots$, we standardly work with the *natural filtration*, denoted by $\mathcal{F}_n \stackrel{\text{def}}{=} \sigma(X_0, X_1, \ldots, X_n)$. This σ-algebra captures exactly the history of the process up to time n. Since the accumulated reward Y_n is a function of the path $X_0 \ldots X_n$, it is strictly $\mathcal{F}_n$-measurable. Moreover, our reward functions are bounded by a constant R,

satisfying the bounded increments condition of Theorem 2.6 whenever the stopping time has finite expectation.

Let $\mathcal{A}$ be a strongly connected Markov chain with state space S, one-step transition probability matrix P and stationary distribution $\pi > \mathbf{0}$. Consider a reward function on the edges $r : E \to [-R, R]$ (alternatively it can be seen as a vector in $[-R, R]^E$) such that the average reward gained in the limit is positive, *i.e.*, $\mu \stackrel{\text{def}}{=} \sum_{e=s \longrightarrow s' \in E} f_e \cdot r(e) > 0$ where $f_e \stackrel{\text{def}}{=} \pi(s) \cdot P(s)(s')$ denotes the long term relative frequency of the edge e. We begin by defining some functions on S^ω. Let $\rho = q_0 q_1 \ldots$ be a generic infinite word.

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

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

$$h \stackrel{\text{def}}{=} \frac{2|S|R}{x_{\min}^{|S|}} \tag{2.1}$$

▶ **Lemma 2.7.** *[BKK14, Theorem 3.4]Let* $u_s \stackrel{def}{=} \sum_{s' \in \mathsf{Succ}(s)} P(s)(s') \cdot r(s, s')$ *be the expected reward gained after taking an edge from state* s*. There exists* $\nu \in [0, h]^S$ *such that*

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

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

▶ **Fact 1.** *Let* ν *be the vector from Lemma 2.7 and* s *be the start state,* i.e.*,* $X_0 = s$*. Then the sequence of random variables given by*

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

is a martingale for all s*.*

Proof. Let $\mathcal{F}_n$ be the natural filtration defined above. We need to show that $\mathcal{E}\left(M_{n+1}^s \mid \mathcal{F}_n\right) = M_n^s$. Recall that $M_n^s = Y_n + \nu(X_n) - n\mu$.

First, we expand the expectation of the next step M_{n+1}^s. Note that $Y_{n+1} = Y_n + r(X_n, X_{n+1})$.

$$\begin{aligned}\mathcal{E}\left(M_{n+1}^s \mid \mathcal{F}_n\right) &= \mathcal{E}(Y_{n+1} + \nu(X_{n+1}) - (n+1)\mu \mid \mathcal{F}_n) \\ &= \mathcal{E}(Y_n + r(X_n, X_{n+1}) + \nu(X_{n+1}) - n\mu - \mu \mid \mathcal{F}_n)\end{aligned}$$

Since Y_n and X_n are $\mathcal{F}_n$-measurable, they act as constants in the conditional expectation. We can pull the known terms out of the expectation:

$$\mathcal{E}\left(M^s_{n+1} \;\middle|\; \mathcal{F}_n\right) = Y_n - n\mu - \mu + \mathcal{E}(r(X_n, X_{n+1}) + \nu(X_{n+1}) \mid X_n)$$

We now evaluate the expectation of the jump, which depends only on the current state X_n. By definition of the transition probabilities:

$$\begin{aligned}\mathcal{E}(r(X_n, X_{n+1}) + \nu(X_{n+1}) \mid X_n) &= \sum_{s' \in \mathsf{Succ}(X_n)} P(X_n)(s') \cdot (r(X_n, s') + \nu(s')) \\ &= \underbrace{\sum_{s'} P(X_n)(s') r(X_n, s')}_{u_{X_n}} + \underbrace{\sum_{s'} P(X_n)(s') \nu(s')}_{(P\nu)(X_n)}\end{aligned}$$

By Lemma 2.7, we have the identity $u + P\nu = \nu + \mathbf{1}\mu$. Applying this to the state X_n yields $u_{X_n} + (P\nu)(X_n) = \nu(X_n) + \mu$. Substituting this back into our derivation for the conditional expectation:

$$\begin{aligned}\mathcal{E}\left(M^s_{n+1} \;\middle|\; \mathcal{F}_n\right) &= Y_n - n\mu - \mu + (\nu(X_n) + \mu) \\ &= Y_n + \nu(X_n) - n\mu \\ &= M^s_n\end{aligned}$$

Thus, the sequence satisfies the martingale property. ◀

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

$$\begin{aligned}&Y^{[a,b]}_0 \stackrel{\text{def}}{=} Y_0 = 0 \\ &Y^{[a,b]}_n \stackrel{\text{def}}{=} \max(a,\, \min(b,\, Y^{[a,b]}_{n-1} + r(X_{n-1}, X_n))) \quad \text{for } n \geq 1\end{aligned}$$

Considering the sequence $Y^{[a,b]}_n$, let $T^{[a,b]}_a$, $T^{[a,b]}_b$ be the functions which denote the first hitting time of the left boundary a and right boundary b respectively. Clearly, $Y^{[a,b]}_n \leq Y_n$ before $T^{[a,b]}_a$, because the only possible difference is that $Y^{[a,b]}_n$ loses something when hitting the right border b. One of the advantages of $Y^{[a,b]}_n$ is that it can be described using only a finite number of bits for any n, because of its boundedness, whereas this is not the case for Y_n. This is useful in situations where such an under-approximation suffices instead of remembering the exact reward gained. However, the bounding of the reward also changes the behavior

of Y_n. For example, the probability of Y_n falling below a infinitely often is zero, *i.e.*, $\mathcal{P}(\mathsf{GF}(Y_n \leq a)) = 0$ for a positive mean payoff $\mu > 0$. The same does not generally hold for $Y_n^{[a,b]}$, which might fall below a infinitely often almost surely.

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

Remark 2.8. Although we denote the functions by $T_a^{[a,b]}$, $T_b^{[a,b]}$ and $Y_n^{[a,b]}$ etc., note that there is an implicit assumption that the initial sum is 0 which lies between a and b. Therefore, the hitting times are actually parametrized by three numbers a, b and x such that $a < x < b$ given by $T_a^{[a,x,b]}$, $T_b^{[a,x,b]}$, $Y_n^{[a,x,b]}$. But we continue to represent with just the boundary points as all the functions are invariant under translation *i.e.*, $T_a^{[a,x,b]} = T_{a'}^{[a',b']}$- where $a' = a - x$ and $b' = b - x$. Moreover, for $a < x_1 < x_2 < b$

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

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

2.2.1 Upper bound on the hitting time in the direction of drift

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

$$T_b \stackrel{\text{def}}{=} \inf\{n \mid Y_n \geq b\}. \tag{2.2}$$

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

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

Now we show some bounds on this expected stopping time.

▶ **Lemma 2.9.** *For all states s and $b > 0$*

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

where h is the constant from Equation (2.1).

Proof. Since $\mathcal{E}_s(T_b) < \infty$ and the martingale M_n^s has bounded step size, ($|M_{n+1}^s - M_n^s| \leq 2R + h$) we can apply Theorem 2.6 to get

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

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

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

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

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

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

Lemma 2.9 shows that, starting from any state, the expected time to hit any upper total reward boundary $b > 0$ is asymptotically linear in b. To get the upper bound for $T_b^{[a,b]}$, observe that $\left(Y_n \leq Y_n^{[a,b]}\right) \mid (T_b \geq n)$ implying $T_b^{[a,b]} \leq T_b$. Hence we obtain the following as a corollary.

▶ **Corollary 2.10.**

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

2.2.2 Bounding the probability of going in the opposite direction of the drift

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

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

$$T_a \stackrel{\text{def}}{=} \inf\{n \mid Y_n \leq a\} \tag{2.3}$$

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

$$\eta \stackrel{\text{def}}{=} \mu + h + R \tag{2.4}$$

$$c \stackrel{\text{def}}{=} e^{\frac{-\mu^2}{2\eta^2}} \tag{2.5}$$

▶ **Lemma 2.11.** *For all $a \leq -h$ and states s we have*

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

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

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

Hence, using the Azuma-Hoeffding inequality, we obtain

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

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

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

◀

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

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

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

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

Intuitively, the behaviour of $Y_n'^{[a,b]}$ is similar to that of $Y_n^{[a,b]}$, except when it hits/exceeds b. Instead of clamping to b, $Y_n'^{[a,b]}$ 'resets' and behaves as if it is

starting from the current state X_n. Let $T_a'^{[a,b]}$ denote the first time that $Y_n'^{[a,b]}$ hits the left bound a, *i.e.*, $T_a'^{[a,b]} \stackrel{\text{def}}{=} \inf\Big\{n \mid Y_n'^{[a,b]} \leq a\Big\}$.

▷ Claim 2.12. For all $n \geq 0$ and $b > 0 > a$ we have $Y_n^{[a,b]} \geq Y_n'^{[a,b]}$. Consequently, $T_a^{[a,b]} \geq T_a'^{[a,b]}$.

Proof. By induction on n.

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

For the induction step let $n > 0$.

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

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

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

◁

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

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

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

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

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

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

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

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

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

▷ Claim 2.13. Let $0 < \delta < 1$ and let c be as in Equation (2.5). By choosing $a \stackrel{\text{def}}{=} \min(-R\lceil \log_c(\delta(1-c))\rceil + R - 1, -h)$ we obtain $\mathcal{P}_s\Big(V_a'^{[a,b]} = 0\Big) \leq \delta$ for any start state s. Moreover, it holds that $\mathcal{P}_s\Big(V_a'^{[a,b]} \geq j+1\Big) \geq (1-\delta)^{j+1}$.

Proof. From our choice of a and Lemma 2.11, it follows that $\mathcal{P}_s^{\mathcal{A}}(T_a < \infty) \leq \frac{c^{\left\lceil \frac{|a|}{R} \right\rceil}}{1-c} \leq \delta$. Consider the event $V_a'^{[a,b]} = 0$ when starting from s. This means any run in this event doesn't hit the reset transitions, which implies that the probability of this event doesn't change when considering Y_n or $Y_n^{[a,b]}$ instead of the sequence $Y_n'^{[a,b]}$.

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

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

$$\mathcal{P}_s\big(V_a'^{[a,b]} = 0\big) = \mathcal{P}_s(T_a < T_b) \leq \mathcal{P}_s^{\mathcal{A}}(T_a < \infty) \leq \delta. \tag{2.6}$$

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

Base case $j = 0$: $\mathcal{P}_s\Big(V_a'^{[a,b]} \geq 1\Big) = \mathcal{P}_s\Big(V_a'^{[a,b]} > 0\Big) = 1 - \mathcal{P}_s\Big(V_a'^{[a,b]} = 0\Big) \geq 1 - \delta$, by Equation (2.6).

Induction step:

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

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

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

◁

Thus, we get that

$$\begin{aligned}\mathcal{E}_s\big(T_a^{[a,b]}\big) \geq \mathcal{E}_s\big(T_a'^{[a,b]}\big) &\geq \alpha + \sum_{j=0}^{\infty} \beta \cdot \mathcal{P}_s\big(V_a'^{[a,b]} \geq j+1\big)\\ &\geq \alpha + \sum_{j=0}^{\infty} \beta \cdot (1-\delta)^{j+1}\\ &= \alpha + \beta \cdot \frac{1-\delta}{\delta}\\ &= \beta \cdot \frac{1}{\delta} + (\alpha - \beta). \qquad (2.7)\end{aligned}$$

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

▶ **Lemma 2.14.** *For $0 < \delta < 1$ and $a = \min(-R\lceil \log_c(\delta(1-c))\rceil + R - 1, -h)$ we have*

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

Proof. Since $b = 1$ and $|a| \geq R(\lceil \log_c(\delta(1-c)) \rceil - 1) + 1$, $\beta = 1$ and $\alpha = \lceil \frac{|a|}{R} \rceil \geq \lceil \log_c(\delta(1-c)) \rceil$. Substituting these values in Equation (2.7) gives the required bound. ◀

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

2.2.3 General Markov Chains

To get a lower bound on $\mathcal{E}_s\left(T_a^{[a,b]}\right)$ when $\mathcal{A}$ is not strongly connected, one has to account for the time spent in transient states. Fortunately, the probability to spend a large amount of time outside a BSCC falls exponentially and this allows the analysis from the previous subsections to carry over with a minimal increase in the size of the interval $[a, b]$ required for general Markov chains. Assume $\texttt{AS}(\texttt{MP}(> 0)) = S$, *i.e.*, every BSCC of $\mathcal{A}$ has positive average mean payoff. Let $C \subseteq S$ denote all the recurrent states. Compute the constants η_G, μ_G, c_G, h_G from Lemma 2.11 for each BSCC G and let η, μ, c, h be the maximum over all BSCC's. Let T_C denote the hitting time of some BSCC. For some positive integer k, let Z_k denote the event $T_C \leq k$. By the tower property, one has

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

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

$$\mathcal{E}_s\big(T_a^{[a,b]}\big) \geq \mathcal{E}_s\big(T_a^{[a,b]} \mid Z_k\big) \cdot \mathcal{P}_s(Z_k). \tag{2.8}$$

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

$$g \stackrel{\text{def}}{=} \exp\left(\frac{-x_{\min}^{|S|}}{|S|}\right) \tag{2.9}$$

▶ **Lemma 2.15.** *Let $y_{\min}$ denote the minimum occurring probability in $\mathcal{A}$ outside every BSCC. Then there exists $0 \leq g < 1$ such that for all $k > |S|$,*

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

Proof. The proof is similar to [BKK14, Lemma 5.1]. Assume $y_{\min} \neq 1$ (the other case is trivial $\mathcal{P}_s\big(\overline{Z_k}\big) = 0$). This implies $y_{\min} \leq \frac{1}{2}$. Let $n = |S|$. From any state s,

there will be a path of length at most $n-1$ to a state in C, $\implies$ for all states s, $\mathcal{P}_s(T_C < n) \geq y_{\min}^{n-1} \geq y_{\min}^n$. Dividing the run into segments of length $n-1$, one gets

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

◀

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

▶ **Lemma 2.16.** *For any $0 < \delta < 1$, choosing $b = 1$ and*

$$a = \min(-R\lceil \log_c(\delta(1-c)) \rceil + R - 1, -h) - k \cdot R$$

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

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

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

$$\begin{aligned}
&\mathcal{E}_s\big(T_a^{[a,b]} \;\big|\; Z_k\big) = \\
&\qquad \sum_{(y,q) \in I} \mathcal{E}_s\Big(T_a^{[a,b]} \;\Big|\; Z_k,\, Y_{T_C}^{[a,b]} = y,\, X_{T_C} = q\Big) \cdot \mathcal{P}_s\Big(Y_{T_C}^{[a,b]} = y, X_{T_C} = q \;\Big|\; Z_k\Big)
\end{aligned}$$

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

$$\begin{aligned}
\mathcal{E}_s\big(T_a^{[a,b]} \;\big|\; E(y,q)\big) &= \mathcal{E}_s(T_C \mid E(y,q)) + \mathcal{E}_q\Big(T_{a-g}^{[a-y,b-g]}\Big) \\
&\geq \mathcal{E}_q\Big(T_{a+k\cdot R}^{[a+k\cdot R, b+k\cdot R]}\Big) \\
&\geq (k+1) \cdot \left(\frac{1}{\delta} - 1 \right) + \lceil \log_c(\delta(1-c)) \rceil \quad ((2.7))
\end{aligned}$$

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

Using Lemmas 2.15 and 2.16, one gets the following result.

▶ **Lemma 2.17.** *For any $k > n$, $0 < \delta < 1$, with $a = \min(-R\lceil \log_c(\delta(1-c))\rceil + R - 1, -h) - k \cdot R$ and $b = 1$*

$$\mathcal{E}_s\big(T_a^{[a,b]}\big) \geq \big(1 - 2 \cdot g^k\big) \cdot \left((k+1) \cdot \left(\frac{1}{\delta} - 1\right) + \lceil \log_c(\delta(1-c))\rceil\right)$$
$$\mathcal{E}_s\Big(T_b^{[a,b]}\Big) \leq |S| + \frac{2}{1-g} + \frac{|a| + b + h + R}{\mu}$$

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

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

Proof. The first result follows from (2.8) and Lemmas 2.15 and 2.16.

To get the upper bound for $T_b^{[a,b]}$ in general case, we simply add the expected time it takes to reach a BSCC and upper bound the time it takes with the worst possible $Y_{T_C}^{[a,b]} (= a)$. Hence, by Lemmas 2.9 and 2.15

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

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

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

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

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

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

$$
\begin{aligned}
&\mathcal{E}_s\big(T_a^{[a,b]} \mid T_a^{[a,b]} < T_C\big) \cdot \mathcal{P}_s\big(T_a^{[a,b]} < T_C\big) \\
&= \sum_{m=k}^{\infty} m \cdot \mathcal{P}_s\big(T_a^{[a,b]} = m \mid T_a^{[a,b]} < T_C\big) \cdot \mathcal{P}_s\big(T_a^{[a,b]} < T_C\big) \\
&= \sum_{m=k}^{\infty} m \cdot \mathcal{P}_s\big(T_a^{[a,b]} = m \cap T_a^{[a,b]} < T_C\big) \\
&\leq \sum_{m=k}^{\infty} m \cdot \mathcal{P}_s(T_C > m) \\
&\leq \sum_{m=k}^{\infty} m \cdot 2 \cdot g^m \quad (\textit{Lemma } 2.15) \\
&\leq \sum_{m=0}^{\infty} m \cdot 2 \cdot g^m \\
&= \frac{2g}{(1-g)^2}
\end{aligned}
$$

and then using above inequality and (2.2.3)

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

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

$$
\begin{aligned}
\mathcal{E}_s\Big(\Big(Y_{T_a^{[a,b]}}\Big)_2 \;\Big|\; T_a^{[a,b]} \geq T_C\Big) \cdot \mathcal{P}_s\big(T_a^{[a,b]} \geq T_C\big) = \\
\mathcal{E}_s\big((Y_{T_C})_2 + (Y_T)_2 \mid T_a^{[a,b]} \geq T_C\big) \cdot \mathcal{P}_s\big(T_a^{[a,b]} \geq T_C\big) \quad (2.10)
\end{aligned}
$$

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

$$
\begin{aligned}
\mathcal{E}_s\big(T_C \mid T_a^{[a,b]} \geq T_C\big) \cdot \mathcal{P}_s\big(T_a^{[a,b]} \geq T_C\big) &\leq \mathcal{E}_s(T_C) \quad \text{Since } T_C \geq 0 \\
&= \sum_{m=0}^{\infty} \mathcal{P}_s(T_C > m) \\
&= |S| + \frac{2 \cdot g^{|S|}}{1-g} \quad (\textit{Lemma } 2.15) \\
&\leq |S| + \frac{2}{1-g}
\end{aligned}
$$

Thus

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

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

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

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

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

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

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

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

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

When $Y_{T_C}^{[a,b]} \geq -k\cdot R$, then $\mathcal{E}_q\Big(T_{a-y}^{[a-y,b-y]}\Big) \geq (k+1)\cdot\big(\frac{1}{\delta}-1\big) + \lceil\log_c(\delta(1-c))\rceil$ using Equation (2.7). Observe that this event subsumes $T_C < k$, so probability of this happening is $\geq 1-2g^k$ by Lemma 2.15. In the other case when $Y_{T_C 1} < -k\cdot R$, a trivial lower bound of 0 for $\mathcal{E}_q\Big(T_{a-y}^{[a-y,b-y]}\Big)$ suffices for our purposes. Since from our assumption, δ is small enough so that $(|S|+1)\cdot\big(\frac{1}{\delta}-1\big) + \lceil\log_c(\delta(1-c))\rceil \geq \frac{h}{\mu_2}$, putting it all together we have

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

So

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

◀

Chapter 3

Approximating the Value of Energy-Parity Objectives in Simple Stochastic Games

3.1 Overview

In this chapter, we look at the problem of approximating the value of states in stochastic games with energy-parity objective. One player (Max) gets a payoff of 1, if they never run out of energy while simultaneously satisfying the qualitative parity condition. In Section 3.3, we state the main results and describe the proof idea. The results are as follows.

1. ε-approximating the value of an energy-parity game can be done in 2-NEXPTIME when the rewards for the energy objective are given in unary.
2. Computing the ε-optimal finite strategies for either player can also be done in 2-NEXPTIME with unary rewards.
3. The above strategies are deterministic and use memory doubly exponential in the size of the game and $\log(1/\varepsilon)$ with unary rewards.

The emphasis on unary rewards is primarily historical, inspired by the work on approximating termination in one-counter stochastic games [BBEK13], where unary updates were considered. The dependency on the size of the game in each of the above results increases by an exponential if the rewards are assumed to be given in binary. To show this, we first compute the limit value of a state

as the initial energy tends to ∞ in Section 3.4. This value turns out to be equal to the value of what we call the '`Gain`' objective. Along with the value, we also compute optimal deterministic strategies for both players. While for Min, the optimal strategy can be chosen memoryless, we show that for Max, finite but exponential memory is both necessary and sufficient when considering deterministic strategies. Once we have these limit values, given ε, we show in Section 3.5, how to compute a number N such that the value of $\texttt{EN}(N) \cap \texttt{EPAR}$ is ε-close to the value of `Gain` from every state s. This is done by first computing this bound in maximizing MDPs(Section 3.5.1) which is then lifted to a bound in general stochastic games(Section 3.5.2). In MDPs, we give an exponential bound on N, when the rewards are in unary. There is an additional exponential blowup in the case of games as the strategies for Max computed in Section 3.4 could be exponential. Finally, Section 3.6 is concerned with computing the strategies and values for the parity objective in the unfolded game until energy level N.

Contributions. The results in this chapter are based on [DM23] published at MFCS 2023.

3.2 Related Work & Contributions

Energy-parity. We consider SSGs with *energy-parity* objectives, where plays need to satisfy both an energy and a parity objective. The parity objective specifies functional correctness, while the energy condition can encode efficiency or risk considerations, e.g., the system should not run out of energy since manually recharging would be costly or risky.

Much work on combined objectives for stochastic systems is restricted to Markov decision processes (MDPs) [CD11a, CD11b, BKN16, MSTW17].

For (stochastic) games, the computational complexity of single objectives is often in $\mathsf{NP} \cap \mathsf{coNP}$, e.g., for parity or mean-payoff objectives [Jur98b]. Multi-objective games can be harder, e.g., satisfying *two different* parity objectives leads to coNP completeness [CHP07].

Stochastic mean-payoff parity games can be solved in $\mathsf{NP} \cap \mathsf{coNP}$ [CDGO14]. However, this does not imply a solution for stochastic energy-parity games, since, unlike in the non-stochastic case [CD10], there is no known reduction from energy-parity to mean-payoff parity in stochastic games. The reduction in [CD10] relies

on the fact that Max has a winning *finite-memory* strategy for energy-parity, which does not generally hold for stochastic games, or even MDPs [MSTW17]. For the same reason, the direct reduction from stochastic energy-parity to ordinary energy games proposed in [CD11a, CD11b] does not work for general energy-parity but only for energy-Büchi; cf. [MSTW17].

Non-stochastic energy-parity games can be solved in $\mathsf{NP} \cap \mathsf{coNP}$ (and even in pseudo-quasi-polynomial time [DJL18]) and Max strategies require only finite (but exponential) memory [CD10].

Stochastic energy-parity games have been studied in [MSTW21], where it was shown that the almost-sure problem is decidable and in $\mathsf{NP} \cap \mathsf{coNP}$. That is, given an initial configuration (control-state plus current energy level), does Max have a strategy to ensure that energy-parity is satisfied with probability 1 against any Min strategy? Unlike in many single-objective games, such an almost-surely winning Max strategy (if it exists) requires infinite memory in general. This holds even in MDPs and for energy-coBüchi objectives [MSTW17].

However, [MSTW21] did not address quantitative questions about energy-parity objectives, such as computing/approximating the value of a given configuration, or the decidability of exact questions like "Is the value of this configuration $\geq k$?" for some constant k (e.g., $k = 1/2$).

The decidability of the latter type of exact question about the energy-parity value is open, but there are strong indications that it is very hard. In fact, even simpler sub-problems are already at least as hard as the *positivity problem for linear recurrence sequences*, which in turn is at least as hard as the *Skolem problem* [EvdPSW03]. (The decidability of these problems has been open for decades; see [OW15] for an overview.) Given an SSG with an energy-parity objective, suppose we remove the parity condition (assume it is always true) and also suppose that Max is passive (does not get to make any decisions). Then we obtain an MDP where the only active player (the Min in the SSG) has a *termination objective*, *i.e.*, to reach a configuration where the energy level is ≤ 0. Exact questions about the value of the termination objective in MDPs are already at least as hard as the positivity problem [Pir21, Section 5.2.3] (see also [PB20, PB23]). Thus exact questions about the energy-parity value in SSGs are also at least as hard as the positivity problem.

Our contributions. Since exact questions about the energy-parity value in SSGs are positivity-hard, we consider the problem of computing approximations of the value. We present an algorithm that, given an SSG $\mathcal{G}$ and error ε, computes ε-close approximations of the energy-parity value of any given configuration in 2-NEXPTIME. Moreover, we show that ε-optimal Max (resp. Min) strategies can be chosen as deterministic and using only finite memory with $\mathcal{O}\big(2\text{-EXP}(\|\mathcal{G}\|) \cdot \log\big(\frac{1}{\varepsilon}\big)\big)$ memory modes. One can understand the idea as a constructive upper bound on the accuracy with which the players need to remember the current energy level in the game. (This is in contrast to the result in [MSTW17] that almost-surely winning Max strategies require infinite memory in general.) Once the upper bound on Max's memory for ε-optimal strategies is established, one might attempt a reduction from energy-parity to mean-payoff parity, along similar lines as for non-stochastic games in [CD10]. However, instead we use a more direct reduction from energy-parity to parity in a derived SSG for our approximation algorithm.

In our constructions we use some auxiliary objectives. Following [MSTW21], these are defined as $\texttt{Gain} \stackrel{\text{def}}{=} \texttt{LimInf}(> -\infty) \cap \texttt{EPAR}$ and $\texttt{Loss} \stackrel{\text{def}}{=} \overline{\texttt{Gain}} = \texttt{LimInf}(= -\infty) \cup \texttt{OPAR}$.

3.3 The Main Result

The following theorem states our main result.

▶ **Theorem 3.1.** *Let $\mathcal{G} = (S, (S_\Box, S_\Diamond, S_\bigcirc), E, P)$ be an SSG with transition rewards in unary assigned by function r and colors assigned to states by function Col. For every state $s \in S$, initial energy level $i \geq 0$ and error margin $\varepsilon > 0$, one can compute*

1. *a rational number v' such that $0 \leq v' - \texttt{val}^{\mathcal{G}}_{\texttt{EN}(i) \cap \texttt{EPAR}}(s) \leq \varepsilon$ in 2-NEXPTIME.*[1]

2. *ε-optimal FDD strategies σ_ε and π_ε for Max and Min, resp., in 2-NEXPTIME. These strategies use $\mathcal{O}\big(2\text{-EXP}(\|\mathcal{G}\|) \cdot \log\big(\frac{1}{\varepsilon}\big)\big)$ memory modes.*

For rewards in binary, the bounds above increase by one exponential.

[1] We write "computing a number v' in 2-NEXPTIME" as a shorthand for the property that questions like $v' \leq c$ for constants c are decidable in 2-NEXPTIME.

Note that the complexity bounds in Theorem 3.1 are independent of the initial energy level i. This is because our algorithm computes ε-optimal strategies for *all* energy levels at once. As we will see in the proof, the construction produces a single strategy structure, based on unfolding the game up to a computed bound N and switching to a limit strategy thereafter that is valid for any starting energy. As a result, the encoding of the initial energy level is immaterial.

We outline the main steps of the proof; details in the following sections. We begin with the observation that $\mathtt{EN}(i) \subseteq \mathtt{EN}(j)$ for $i \leq j$, and thus for all states s we have $\mathtt{val}^{\mathcal{G}}_{\mathtt{EN}(i)\cap\mathtt{EPAR}}(s) \leq \mathtt{val}^{\mathcal{G}}_{\mathtt{EN}(j)\cap\mathtt{EPAR}}(s) \leq 1$. So $\lim_{n\to\infty} \mathtt{val}^{\mathcal{G}}_{\mathtt{EN}(n)\cap\mathtt{EPAR}}(s)$ exists. We define

$$\mathtt{Lval}^{\mathcal{G}}(s) \stackrel{\text{def}}{=} \lim_{n\to\infty} \mathtt{val}^{\mathcal{G}}_{\mathtt{EN}(n)\cap\mathtt{EPAR}}(s). \tag{3.1}$$

We will see that $\mathtt{Lval}^{\mathcal{G}}(s)$ and $\mathtt{val}^{\mathcal{G}}_{\mathtt{Gain}}(s)$ are in fact equal (a consequence of Lemma 3.8) and $\mathtt{val}^{\mathcal{G}}_{\mathtt{Gain}}(s)$ can be computed in nondeterministic polynomial time (Theorem 3.4). Intuitively, for high energy levels, the precise energy level does not matter much for the value.

The main steps of the approximation algorithm are as follows.

1. Compute FDD strategies $\sigma^*(s)$ that are optimal maximizing for the objective $\mathtt{Gain}$ starting from state s in $\mathcal{G}$. Compute an MD strategy π^* that is uniformly optimal minimizing for the objective $\mathtt{Gain}$. Compute the value $\mathtt{val}^{\mathcal{G}}_{\mathtt{Gain}}(s)$ for every $s \in S$. See Section 3.4.

2. Compute a natural number N such that for all $s \in S$ and all $i \geq N$ we have

$$0 \leq \mathtt{val}^{\mathcal{G}}_{\mathtt{Gain}}(s) - \mathtt{val}^{\mathcal{G}}_{\mathtt{EN}(i)\cap\mathtt{EPAR}}(s) \leq \varepsilon.$$

   N will be doubly exponential. See Section 3.5.

3. Consider the finite-state parity game $\mathcal{G}'$ derived from $\mathcal{G}$ by encoding the energy level up-to N into the states, *i.e.*, the states of $\mathcal{G}'$ are of the form (s,k) for $s \in S$ and $0 \leq k \leq N$, and colors are inherited from s. Moreover, we add gadgets that ensure that states (s,N) at the upper end win with probability $\mathtt{val}^{\mathcal{G}}_{\mathtt{Gain}}(s)$ and states $(s,0)$ at the lower end lose. By the previous item, $\mathtt{val}^{\mathcal{G}}_{\mathtt{Gain}}(s)$ is ε-close to $\mathtt{val}^{\mathcal{G}}_{\mathtt{EN}(N)\cap\mathtt{EPAR}}(s)$. Thus, for $k < N$ we can ε-approximate the value $v = \mathtt{val}^{\mathcal{G}}_{\mathtt{EN}(k)\cap\mathtt{EPAR}}(s)$ by $v' \stackrel{\text{def}}{=} \mathtt{val}^{\mathcal{G}'}_{\mathtt{EPAR}}((s,k))$. If $k \geq N$ we can ε-approximate v by $v' \stackrel{\text{def}}{=} \mathtt{val}^{\mathcal{G}}_{\mathtt{Gain}}(s)$.

Moreover, we obtain ε-optimal FDD strategies σ_ε for Max (resp. π_ε for Min) for $\texttt{EN}(k) \cap \texttt{EPAR}$ in $\mathcal{G}$. Let $\hat{\sigma}$ (resp. $\hat{\pi}$) be optimal MD strategies for Max (resp. Min) for the objective $\texttt{EPAR}$ in $\mathcal{G}'$. Then σ_ε plays as follows. While the current energy level j (k plus the sum of the rewards so far) stays $< N$, then, at any state s', play like $\hat{\sigma}$ at state (s', j) in $\mathcal{G}'$. Once the energy level reaches a value $\geq N$ at some state s' for the first time, then play like $\sigma^*(s')$ forever. Similarly, π_ε plays as follows. While the current energy level j (k plus the sum of the rewards so far) stays $< N$, then, at any state s', play like $\hat{\pi}$ at state (s', j) in $\mathcal{G}'$. Once the energy level reaches a value $\geq N$ (at any state) for the first time, then play like π^* forever. See Section 3.6.

As a technical tool, we sometimes consider the dual of a game $\mathcal{G}$ (resp. the dual maximizing MDP of some minimizing MDP). We denote the dual of $\mathcal{G}$ by $\mathcal{G}^d \stackrel{\text{def}}{=} \left(S', (S'_\Box, S'_\Diamond, S'_\bigcirc), E', P'\right)$ with the complement objective $\overline{\texttt{EN}(k) \cap \texttt{EPAR}} = \texttt{Term}(k) \cup \texttt{OPAR}$, where $\mathcal{G}^d$ is simply the game with the roles of Max and Min reversed, *i.e.*,

$$S' = S \qquad S'_\Box = S_\Diamond \qquad S'_\Diamond = S_\Box \qquad S'_\bigcirc = S_\bigcirc \qquad E' = E \qquad P' = P$$

Hence $\Sigma_{\mathcal{G}^d} = \Pi_\mathcal{G}$ and $\Pi_{\mathcal{G}^d} = \Sigma_\mathcal{G}$. It is easy to see that for any objective $\texttt{O}$ and start state s

1. $\texttt{val}_{\texttt{O}}^{\mathcal{G}}(s) + \texttt{val}_{\overline{\texttt{O}}}^{\mathcal{G}^d}(s) = 1$.
2. σ is ε-optimal maximizing for $\texttt{O}$ in $\mathcal{G}$ iff it is ε-optimal minimizing for $\overline{\texttt{O}}$ in $\mathcal{G}^d$.
3. π is ε-optimal minimizing for $\texttt{O}$ in $\mathcal{G}$ iff it is ε-optimal maximizing for $\overline{\texttt{O}}$ in $\mathcal{G}^d$.

So approximating the value of $\texttt{EN}(k) \cap \texttt{EPAR}$ in $\mathcal{G}$ can be reduced in linear time to approximating the value of $\texttt{Term}(k) \cup \texttt{OPAR}$ in $\mathcal{G}^d$.

3.4 Computing $\texttt{val}_{\texttt{Gain}}^{\mathcal{G}}(s)$

Given an SSG $\mathcal{G} = (S, (S_\Box, S_\Diamond, S_\bigcirc), E, P)$ and a start state s, we will show how to compute $\texttt{val}_{\texttt{Gain}}^{\mathcal{G}}(s)$ and the optimal strategies for both players.

We start with the case of maximizing MDPs. The following lemma summarizes some previous results ([MSTW21, Lemmas 30,16], [MSTW17, Lemma 26], [GOP11, Proposition 4]).

▶ **Lemma 3.2.** *Let $\mathcal{M}$ be a maximizing MDP.*

1. $\mathtt{Lval}^{\mathcal{M}}(s) = \mathtt{val}^{\mathcal{M}}_{\mathtt{Gain}}(s)$ *for all states* $s \in S$.
2. *Optimal strategies for* `Gain` *in* $\mathcal{M}$ *exist and can be chosen FDD, with at most* $\mathcal{O}\big(\exp\big(\|\mathcal{M}\|^{\mathcal{O}(1)}\big)\big)$ *memory modes, and exponential memory is also necessary.*
3. *For any state* $s \in S$, $\mathtt{Lval}^{\mathcal{M}}(s)$ *is rational and can be computed in* $\mathcal{O}(\|\mathcal{M}\|^8)$ *deterministic polynomial time if rewards are in unary, and in* NP *and* coNP *if rewards are in binary.*

Proof. Item 1. holds by [MSTW21, Lemma 30].

Towards Item 2., we follow the proof of [MSTW21, Lemma 16]. Since $\mathtt{Gain} = \mathtt{LimInf}(> -\infty) \cap \mathtt{EPAR}$ is shift-invariant, there exist optimal strategies by [GH10]. By [MSTW17, Theorem 18] and Item 1., an optimal strategy for `Gain` can be constructed as follows. Let $A \stackrel{\text{def}}{=} \bigcup_{k\in\mathbb{N}} \mathsf{AS}\left(\mathtt{ST}k\right) \cap \mathtt{EPAR}\right)$ and $B \stackrel{\text{def}}{=} \mathsf{AS}\left(\mathtt{LimInf}(=\infty) \cap \mathtt{EPAR}\right)$ be the subsets of states from which there exist almost surely winning strategies for the objectives $\mathtt{ST}k) \cap \mathtt{EPAR}$ and $\mathtt{LimInf}(=\infty) \cap \mathtt{EPAR}$, respectively. By [MSTW17, Theorem 8], we can restrict the values k in the definition of A by some $k' = \mathcal{O}(|S| \cdot R)$, *i.e.*, $A = \bigcup_{k\leq k'} \mathsf{AS}\left(\mathtt{ST}k\right) \cap \mathtt{EPAR}\right)$. An optimal strategy σ for `Gain` works in two phases. First it plays an optimal strategy σ_R towards reaching the set $A \cup B$, where σ_R can be chosen MD by Remark 2.1. Then, upon reaching A (resp. B), it plays an almost surely winning strategy σ_A for the objective $\mathtt{ST}k) \cap \mathtt{EPAR}$ (resp. σ_B for the objective $\mathtt{LimInf}(=\infty) \cap \mathtt{EPAR}$). By [MSTW17, Theorem 8], the strategy σ_A requires $\mathcal{O}(k \cdot |S|)$ memory modes for a given k and thus at most $\mathcal{O}(|S|^2 \cdot R)$, since we can assume that $k \leq k'$. Towards the strategy σ_B, we first observe that in finite MDPs a strategy is almost-surely winning for $\mathtt{LimInf}(=\infty) \cap \mathtt{EPAR}$ iff it is almost-surely winning for $\mathtt{MP}>0 \cap \mathtt{EPAR}$. By [GOP11, Proposition 4], there exist optimal deterministic strategies for $\mathtt{MP}>0 \cap \mathtt{EPAR}$ that use exponential memory, *i.e.*, $\mathcal{O}(\exp(\|\mathcal{M}\|^{\mathcal{O}(1)}))$ memory modes. The memory required for σ_B exceeds that of σ_R and σ_A (even when R is given in binary), and the one extra memory mode to record the switch from σ_R to σ_A (resp. σ_B) is negligible in comparison. Thus, we can conclude that σ uses $\mathcal{O}(\exp(\|\mathcal{M}\|^{\mathcal{O}(1)}))$ memory modes. [GOP11, Fig. 1 and Prop. 4] shows that exponential memory is necessary.

Towards Item 3., let $d \stackrel{\text{def}}{=} |Col(S)|$ be the number of priorities in the parity condition. By [MSTW17, Lemma 26], for each $s \in S$, $\mathtt{Lval}^{\mathcal{M}}(s)$ is rational and

can be computed in deterministic time $\tilde{\mathcal{O}}(|E| \cdot d \cdot |S|^4 \cdot R + d \cdot |S|^{3.5} \cdot (\|P\| + \|r\|)^2)$ (and still in NP and coNP if R is given in binary). So $\texttt{Lval}^{\mathcal{M}}(s)$ can be computed in $\mathcal{O}(\|\mathcal{M}\|^8)$ deterministic polynomial time if weights are given in unary, and in NP and coNP if weights are given in binary. ◀

In order to extend Lemma 3.2 from MDPs to games, we need to link the value of a state s in the game $\mathcal{G}$ to its value in the induced maximizing (resp. minimizing) MDP when fixing a possibly optimal Min (resp. Max) FR strategy π (resp. σ). The following lemma is straightforward from the definition of induced MDP and the fact that infimum (resp. supremum) is at most (resp. at least) any number in the set.

▶ **Lemma 3.3.** *For every SSG $\mathcal{G}$, objective $\texttt{O}$[2] and Min (resp. Max) FDD strategy $\pi = (\mathsf{M}, \mathsf{m}_0, \mathsf{upd}, \mathsf{nxt})$ (resp. σ), from state s we get $\texttt{val}_{\texttt{O}}^{\mathcal{G}^\sigma}((\mathsf{m}_0, s)) \leq \texttt{val}_{\texttt{O}}^{\mathcal{G}}(s) \leq \texttt{val}_{\texttt{O}}^{\mathcal{G}_\pi}((\mathsf{m}_0, s))$ and equality holds if π (resp. σ) is optimal from state s.*

▶ **Theorem 3.4.** *Consider a SSG $\mathcal{G} = (S, (S_\Box, S_\Diamond, S_\bigcirc), E, P)$ with the* Gain *objective.*

1. *Optimal Min strategies exist and can be chosen uniform MD.*
2. $\texttt{val}_{\texttt{Gain}}^{\mathcal{G}}(s)$ *is rational and questions about it,* i.e., $\texttt{val}_{\texttt{Gain}}^{\mathcal{G}}(s) \leq c$ *for constants c, are decidable in* NP.
3. *Optimal Max strategies exist and can be chosen FDD, with $\mathcal{O}\big(\exp\big(\|\mathcal{G}\|^{\mathcal{O}(1)}\big)\big)$ memory modes. Moreover, exponential memory is also necessary.*

Proof. Towards Item 1., observe that since both the objectives $\texttt{LimInf}(= -\infty)$ and OPAR are shift-invariant and submixing, so is their union, *i.e.*, $\overline{\texttt{Gain}}$ is shift-invariant and submixing. Hence, by [GK23, Theorem 1.1], an optimal MD strategy π_s^* for Min exists from any state $s \in S$. Since S is finite and Gain is shift-invariant, we can also obtain a uniformly optimal MD strategy π^*, *i.e.*, π^* is optimal from every state.

Towards Item 2., consider the maximizing MDP $\mathcal{G}_{\pi^*}$ obtained from $\mathcal{G}$ by fixing π^*. Since π^* is MD, the states of $\mathcal{G}_{\pi^*}$ are the same as the states as $\mathcal{G}$. Since π^* is optimal for Min from every state s, we obtain that $\texttt{val}_{\texttt{Gain}}^{\mathcal{G}}(s) = \texttt{val}_{\texttt{Gain}}^{\mathcal{G}_{\pi^*}}(s)$ for every state s by Lemma 3.3. By Lemma 3.2, the latter is rational and can be

[2] Technically, the objective should change to accommodate for the change in the state space of the induced game but we denote it by the same symbol

computed in polynomial time for weights in unary (resp. in NP and coNP for weights in binary). Thus, by guessing π^*, we can decide questions $\mathtt{val}^{\mathcal{G}}_{\mathtt{Gain}}(s) \leq c$ in NP.

Towards Item 3., we again use the property that $\overline{\mathtt{Gain}}$ is shift-invariant and submixing (see above). By [MSTW21, Theorem 6, Def. 24], optimal FDD Max strategies for $\mathtt{Gain}$ in an SSG require only $|S_\diamond| \cdot \lceil \log(|E|) \rceil$ many extra bits of memory above the memory required for optimal Max strategies in any derived MDP $\mathcal{M}$ where Min's choices are fixed. Hence, by Lemma 3.2, one can obtain optimal FDD Max strategies in $\mathcal{G}$ that use at most $2^{|S_\diamond| \cdot \lceil \log(|E|) \rceil} \cdot \mathcal{O}(\exp(\|\mathcal{M}\|^{\mathcal{O}(1)})) = \mathcal{O}(\exp(\|\mathcal{G}\|^{\mathcal{O}(1)}))$ memory modes. The corresponding exponential lower bound on the memory holds already for MDPs by Lemma 3.2. ◀

3.5 Computing the Upper Bound N

We show how to compute the upper bound N, up-to which Max needs to remember the energy level, for any given error margin $\varepsilon > 0$. Similarly as in Section 3.4, we first solve the problem for maximizing MDPs and then extend the solution to SSGs.

3.5.1 Computing N for maximizing MDPs

Given a maximizing MDP $\mathcal{M} = (S, S_\Box, S_\bigcirc, E, P)$ and $\varepsilon > 0$, we will compute an $N \in \mathbb{N}$ such that for all $s \in S$ and all $j \geq N$

$$0 \leq \mathtt{val}^{\mathcal{M}}_{\mathtt{Term}(j) \cup \mathtt{OPAR}}(s) - \mathtt{val}^{\mathcal{M}}_{\mathtt{Loss}}(s) \leq \varepsilon.$$

Recall that $\mathtt{Loss} = \mathtt{LimInf}(= -\infty) \cup \mathtt{OPAR}$. We now define the sets of states $W_0 \stackrel{\text{def}}{=} \mathsf{AS}(\mathtt{Loss})$, $W_1 \stackrel{\text{def}}{=} \mathsf{AS}(\mathtt{LimInf}(= -\infty))$ and $W_2 \stackrel{\text{def}}{=} \mathsf{AS}(\mathtt{OPAR})$. By Remark 2.1, there exist optimal MD strategies for $\mathtt{LimInf}(= -\infty)$ and $\mathtt{OPAR}$. Since $\mathtt{Loss}$ is shift-invariant and submixing, there exists an optimal MD strategy for it by [GK23, Theorem 1.1].

▶ **Lemma 3.5.** *For every state s in the MDP $\mathcal{M}$ we have*

1. $W_1 \cup W_2 \subseteq W_0$
2. $\mathtt{val}_{\mathsf{F}\, W_0}(s) \leq \mathtt{val}_{\mathtt{Loss}}(s)$
3. $\mathtt{val}_{\mathtt{OPAR} \cap \overline{\mathsf{F}\, W_2}}(s) = 0$

4. *for every initial energy level $j \geq 0$*

$$\mathtt{val}_{(\mathtt{Term}(j) \cup \mathtt{OPAR}) \cap \mathsf{F}\, W_0}(s) = \mathtt{val}_{\mathsf{F}\, W_0}(s) \tag{3.2}$$

$$\mathtt{val}_{\mathtt{Loss}}(s) \leq \mathtt{val}_{\mathtt{Term}(j) \cup \mathtt{OPAR}}(s) \leq \mathtt{val}_{\mathtt{Loss}}(s) + \sup_{\sigma} \mathcal{P}_{\sigma,s}\big(\mathtt{Term}(j) \cap \overline{\mathsf{F}\, W_1}\big) \tag{3.3}$$

Proof.

1. This follows directly from the definitions of W_0, W_1, W_2.

2. Let σ' be an optimal MD strategy for $\mathsf{F}\, W_0$ from s and σ'' be an almost surely winning MD strategy for $\mathtt{Loss}$ from any state in W_0. Let σ be the strategy that plays σ' until reaching W_0 and then switches to σ''. We have $\mathtt{val}_{\mathtt{Loss}}(s) \geq \mathcal{P}_{\sigma,s}(\mathtt{Loss}) \geq \mathcal{P}_{\sigma',s}(\mathsf{F}\, W_0) = \mathtt{val}_{\mathsf{F}\, W_0}(s)$.

3. For $s \in W_2$ the statement is obvious. So let $s \notin W_2$ and consider the modified MDP $\mathcal{M}' = \big(S', S'_\Box, S'_\bigcirc, E', P'\big)$ where all states in W_2 are collapsed into a losing sink. I.e., $S' \stackrel{\text{def}}{=} (S \setminus W_2) \uplus \{\mathit{trap}\}$, with *trap* a new random sink state having color 0 (thus losing for objective $\mathtt{OPAR}$), E' contains all of $(E \cap \{(S \setminus W_2) \times (S \setminus W_2)\} \cup (\mathit{trap}, \mathit{trap}))$ and all transitions to W_2 are redirected to *trap* and P' is derived accordingly from P. Then $\mathtt{val}^{\mathcal{M}'}_{\mathtt{OPAR}}(\hat{s}) = \mathtt{val}^{\mathcal{M}}_{\mathtt{OPAR} \cap \overline{\mathsf{F} W_2}}(\hat{s})$ for all states $\hat{s} \in S \setminus W_2$. Towards a contradiction, assume that $\mathtt{val}^{\mathcal{M}}_{\mathtt{OPAR} \cap \overline{\mathsf{F} W_2}}(s) > 0$. Hence $\mathtt{val}^{\mathcal{M}'}_{\mathtt{OPAR}}(s) > 0$. Then, by [GH10, Theorem 3.2], there exists a state $s' \in S'$ such that $\mathtt{val}^{\mathcal{M}'}_{\mathtt{OPAR}}(s') = 1$, and it is easy to see that $s' \neq \mathit{trap}$ and thus $s' \in S \setminus W_2$. But this implies that $\mathtt{val}^{\mathcal{M}}_{\mathtt{OPAR}}(s') = 1$ and thus $s' \in W_2$, a contradiction.

4. Let $\mathtt{O} \stackrel{\text{def}}{=} \mathtt{Term}(j) \cup \mathtt{OPAR}$. For (3.2), the first inequality $\mathtt{val}_{\mathtt{O} \cap \mathsf{F} W_0}(s) \leq \mathtt{val}_{\mathsf{F} W_0}(s)$ is trivial, since $\mathtt{O} \cap \mathsf{F}\, W_0 \subseteq \mathsf{F} W_0$. To show the reverse inequality, consider the strategy σ that first plays like an optimal MD strategy σ' for the objective $\mathsf{F}\, W_0$ and after reaching W_0 switches to an almost surely winning MD strategy σ'' for the objective $\mathtt{Loss}$. Then $\mathtt{val}_{\mathtt{O} \cap \mathsf{F} W_0}(s) \geq \mathcal{P}_{\sigma,s}(\mathtt{O} \cap \mathsf{F}\, W_0) \geq \mathcal{P}_{\sigma,s}(\mathtt{Loss} \cap \mathsf{F}\, W_0) = \mathcal{P}_{\sigma',s}(\mathsf{F}\, W_0) = \mathtt{val}_{\mathsf{F} W_0}(s)$, where the second inequality is due to $\mathtt{LimInf}(= -\infty) \subseteq \mathtt{Term}(j)$.

   For (3.3), the first inequality is again due to the fact that $\mathtt{LimInf}(= -\infty) \subseteq \mathtt{Term}(j)$ for all $j \geq 0$. Towards the second inequality of Equation (3.3) we

have

$$
\begin{aligned}
&\mathtt{val}_{\mathtt{0}}(s)\\
&= \sup_{\sigma} \mathcal{P}_{\sigma,s}(\mathtt{0})\\
&= \sup_{\sigma} \big(\mathcal{P}_{\sigma,s}(\mathtt{0} \cap \mathsf{F}\, W_0) + \mathcal{P}_{\sigma,s}\big(\mathtt{0} \cap \overline{\mathsf{F}\, W_0}\big)\big) && \text{(Law of total probability)}\\
&\leq \sup_{\sigma} \mathcal{P}_{\sigma,s}(\mathtt{0} \cap \mathsf{F}\, W_0) + \sup_{\sigma} \mathcal{P}_{\sigma,s}\big(\mathtt{0} \cap \overline{\mathsf{F}\, W_0}\big) && (\sup(f+g) \leq \sup f + \sup g)\\
&= \sup_{\sigma} \mathcal{P}_{\sigma,s}(\mathsf{F}\, W_0) + \sup_{\sigma} \mathcal{P}_{\sigma,s}\big(\mathtt{0} \cap \overline{\mathsf{F}\, W_0}\big) && (\textit{Equation } (3.2))\\
&\leq \mathtt{val}_{\mathtt{Loss}}(s) + \sup_{\sigma} \mathcal{P}_{\sigma,s}\big(\mathtt{0} \cap \overline{\mathsf{F}\, W_0}\big) && (\textit{Item } 2.)
\end{aligned}
$$

We can upper-bound the second summand above as follows.

$$
\begin{aligned}
&\sup_{\sigma} \mathcal{P}_{\sigma,s}\big(\mathtt{0} \cap \overline{\mathsf{F}\, W_0}\big)\\
&= \sup_{\sigma} \mathcal{P}_{\sigma,s}\big((\mathtt{Term}(j) \cup \mathtt{OPAR}) \cap \overline{\mathsf{F}\, W_0}\big)\\
&\leq \sup_{\sigma} \mathcal{P}_{\sigma,s}\big(\mathtt{Term}(j) \cap \overline{\mathsf{F}\, W_0}\big) + \sup_{\sigma} \mathcal{P}_{\sigma,s}\big(\mathtt{OPAR} \cap \overline{\mathsf{F}\, W_0}\big) && \text{(Union bound)}\\
&\leq \sup_{\sigma} \mathcal{P}_{\sigma,s}\big(\mathtt{Term}(j) \cap \overline{\mathsf{F}\, W_1}\big) + \sup_{\sigma} \mathcal{P}_{\sigma,s}\big(\mathtt{OPAR} \cap \overline{\mathsf{F}\, W_2}\big) && (\textit{Item } 1.)\\
&= \sup_{\sigma} \mathcal{P}_{\sigma,s}\big(\mathtt{Term}(j) \cap \overline{\mathsf{F}\, W_1}\big) && (\textit{Item } 3.) \quad \blacktriangleleft
\end{aligned}
$$

We show that the term $\sup_\sigma \mathcal{P}_{\sigma,s}\big(\mathtt{Term}(j) \cap \overline{\mathsf{F}\, W_1}\big)$ in Equation (3.3) can be made arbitrarily small for large j. To this end, we use [BBEK13, Lemma 3.9] (adapted to our notation).

▶ **Lemma 3.6.** *[BBEK13, Lemma 3.9 and Claim 6] Let $\mathcal{M} = (S, S_\Box, S_\bigcirc, E, P)$ be a maximizing finite MDP with rewards in unary and $W_1 \stackrel{def}{=} \mathsf{AS}\,(\mathtt{LimInf}(= -\infty))$. One can compute, in polynomial time, a rational constant $c < 1$, and an integer $h \geq 0$ such that for all $j \geq h$ and $s \in S$*

$$\sup_{\sigma} \mathcal{P}_{\sigma,s}\big(\mathtt{Term}(j) \cap \overline{\mathsf{F}\, W_1}\big) \leq \frac{c^j}{1-c}.$$

Moreover, $1/(1-c) \in \mathcal{O}\left(\exp(\|\mathcal{M}\|^{\mathcal{O}(1)})\right)$ and $h \in \mathcal{O}(\exp\big(\|\mathcal{M}\|^{\mathcal{O}(1)}\big))$.

▶ **Lemma 3.7.** *Consider a maximizing MDP $\mathcal{M} = (S, S_\Box, S_\bigcirc, E, P)$, $\varepsilon > 0$ and the constants c, h from Lemma 3.6. For rewards in unary and $i \geq N$ we have $\mathtt{val}^{\mathcal{M}}_{\mathtt{Term}(i) \cup \mathtt{OPAR}}(s) - \mathtt{val}^{\mathcal{M}}_{\mathtt{Loss}}(s) \leq \varepsilon$ where $N \stackrel{def}{=} \max(h, \lceil \log_c(\varepsilon \cdot (1-c)) \rceil) \in \mathcal{O}\left(\exp(\|\mathcal{M}\|^{\mathcal{O}(1)}) \cdot \log(1/\varepsilon)\right)$.*

For rewards in binary we have $N \in \mathcal{O}\left(\exp(\exp(\|\mathcal{M}\|^{\mathcal{O}(1)})) \cdot \log(1/\varepsilon)\right)$, i.e., the size of N increases by one exponential.

Proof sketch. For rewards in unary, the result follows from Lemma 3.5(Equation (3.3)) and Lemma 3.6. For rewards in binary, the constants increase by one exponential via encoding binary rewards into unary rewards in a modified MDP.

Proof. By Lemma 3.5(Equation (3.3)) and Lemma 3.6, we have

$$\mathtt{val}^{\mathcal{M}}_{\mathtt{Term}(i)\cup\mathtt{OPAR}}(s) - \mathtt{val}^{\mathcal{M}}_{\mathtt{Loss}}(s) \leq \sup_{\sigma} \mathcal{P}_{\sigma,s}\big(\mathtt{Term}(i) \cap \overline{\mathsf{F}\, W_1}\big) \leq \frac{c^i}{1-c}$$

for all $i \geq h$ and $s \in S$. To obtain a bound $N \geq h$ with $\frac{c^N}{1-c} \leq \varepsilon$, it suffices to choose

$$N \stackrel{\text{def}}{=} \max(h, \lceil \log_c(\varepsilon \cdot (1-c)) \rceil).$$

We observe that $\log_c(\varepsilon \cdot (1-c)) = -\ln(\varepsilon \cdot (1-c)) \cdot (-\ln(c)^{-1})$.
However, $-\ln(c) = -\ln(1-(1-c)) \geq (1-c)$. Thus $\log_c(\varepsilon \cdot (1-c)) \leq \ln\big(\frac{1}{\varepsilon} \cdot \frac{1}{1-c}\big) \cdot \frac{1}{1-c}$. For rewards in unary, by Lemma 3.6, we have $1/(1-c) \in \mathcal{O}\left(\exp(\|\mathcal{M}\|^{\mathcal{O}(1)})\right)$ and h is only $\mathcal{O}(\exp\big(\|\mathcal{M}\|^{\mathcal{O}(1)}\big))$. Thus $N \in \mathcal{O}\left(\exp(\|\mathcal{M}\|^{\mathcal{O}(1)}) \cdot \log(1/\varepsilon)\right)$.

Now consider the case where rewards are given in binary. Following the proof of [BBEK13, Lemma 3.9], the bounds are derived from the size of solutions of the constructed linear program. While the MDPs in [BBEK13] only consider unary rewards from $\{-1, 0, 1\}$, one can extend it to the case where the rewards come from the set $\{-R, \ldots, 0, \ldots, R\}$ in a natural way. This affects the complexity of the above computed constants and thereby size of N. More precisely, the proof of Lemma 3.6 can be split into three steps. Firstly, given an MDP $\mathcal{M}$ construct a new "rising" MDP $\mathcal{M}'$. Then from this derived $\mathcal{M}'$, construct a linear program. From the solutions of constructed LP, compute the required c and h. We evaluate the effect of having non-unary rewards in each of these steps.

When rewards are given in unary, the resulting $\mathcal{M}'$ has overall size $\|\mathcal{M}'\| \leq 10\|\mathcal{M}\|^4$. More exactly, $|S'| \leq 10 * |S|^3 * (|S| + |E|)$ and similarly for $|E'|$. When the rewards are given in binary, the construction results in an additional R^2 factor. So the resulting $\mathcal{M}'$ is pseudo-polynomially big when compared to $\mathcal{M}$ in our case.

The constructed LP (cf. [BBEK13, Fig.1]) has $S' + 2$ variables (z_s for each state, x for the mean payoff and ξ for converting the constraint $x > 0$ to $x \geq \xi$). Moreover all variables can be assumed non-negative. The number of constraints is bounded by $E' + 1$. Furthermore all the constants appearing in the constraints are either constants in the original MDP $\mathcal{M}$ or 1 or 0.

Finally, from an optimal solution of the LP (z_s, x, ξ) one can compute $\exp\left(\frac{-x^2}{2\cdot(z_{\max}+x+R)^2}\right)$ and to get c, then take a rational over-approximation

and also take h as $\lceil z_{\max} \rceil$ where $z_{\max} \stackrel{\text{def}}{=} \max_{s \in S'} z_s - \min_{s \in S'} z_s$. The only difference compared to the unary rewards case here is that the one-step change of the submartingale is bounded by $z_{\max} + x + R$ instead of $z_{\max} + x + 1$.

From the complexity point of view, both the construction of the LP and the computation from its optimal solutions aren't affected by changes in the rewards, *i.e.*, the previous bounds for c, h and N in terms of $\|\mathcal{M}'\|$ still hold. In particular, $c \in \mathcal{O}\left(\exp\left(1/2^{\|\mathcal{M}'\|^{\mathcal{O}(1)}}\right)\right)$, $h \in \mathcal{O}(\exp(\|\mathcal{M}'\|^{\mathcal{O}(1)}))$ and thus $N \in \mathcal{O}\big(\exp(\|\mathcal{M}'\|^{\mathcal{O}(1)}) \cdot \log(1/\varepsilon)\big)$ by [BBEK13, Claim 6].

While previously, $\mathcal{M}'$ is only polynomially larger than $\mathcal{M}$, introducing binary rewards blows up the construction (cf. [BBEK13, Appendix A.2]). As a result we have that $\|\mathcal{M}'\| \in \mathcal{O}\left(2^{\|\mathcal{M}\|^{\mathcal{O}(1)}}\right)$. Therefore N can be doubly exponential in the size of the original MDP $\mathcal{M}$, *i.e.*, $N \in \mathcal{O}\left(\exp(\exp(\|\mathcal{M}\|^{\mathcal{O}(1)})) \cdot \log(1/\varepsilon)\right)$. ◀

3.5.2 Computing N for SSGs

In order to compute the bound N for an SSG $\mathcal{G}$, we first consider bounds $N(s)$ for individual states s and then take their maximum. Given a state s, we can use Theorem 3.4(Item 3.) to obtain an optimal FDD strategy (with $\mathcal{O}(\exp(\|\mathcal{G}\|^{\mathcal{O}(1)}))$ memory modes) $\sigma^*(s) = (\mathsf{M}, \mathsf{m}_0, \mathsf{upd}, \mathsf{nxt})$ for Max from state s w.r.t. the $\mathtt{Gain}$ objective. Theorem 3.4(Item 1.) yields a uniform MD strategy π^* that is optimal for Min from all states s w.r.t. the $\mathtt{Gain}$ objective.

▶ **Lemma 3.8.** *Given an SSG $\mathcal{G} = (S, (S_\Box, S_\Diamond, S_\bigcirc), E, P)$ and $\varepsilon > 0$, we can compute a number $N \in \mathbb{N}$ such that for all $i \geq N$ and states $s \in S$ we have*

$$\mathtt{val}^{\mathcal{G}}_{\mathtt{EN}(i) \cap \mathtt{EPAR}}(s) - \varepsilon \leq \mathtt{val}^{\mathcal{G}}_{\mathtt{Gain}}(s) - \varepsilon \leq \inf_{\pi} \mathcal{P}^{\mathcal{G}}_{\sigma^*(s),\pi,s}(\mathtt{EN}(i) \cap \mathtt{EPAR}) \leq \mathtt{val}^{\mathcal{G}}_{\mathtt{EN}(i) \cap \mathtt{EPAR}}(s) \tag{3.4}$$

i.e., *$\sigma^*(s)$ is ε-optimal for Max for $\mathtt{EN}(i) \cap \mathtt{EPAR}$ for all $i \geq N$. In particular,* $0 \leq \mathtt{val}^{\mathcal{G}}_{\mathtt{Gain}}(s) - \mathtt{val}^{\mathcal{G}}_{\mathtt{EN}(i) \cap \mathtt{EPAR}}(s) \leq \varepsilon$.

Moreover, π^* is ε-optimal for Min from any state s for $i \geq N$.

$$\sup_{\sigma} \mathcal{P}^{\mathcal{G}}_{\sigma,\pi^*,s}(\mathtt{EN}(i) \cap \mathtt{EPAR}) \leq \sup_{\sigma} \mathcal{P}^{\mathcal{G}}_{\sigma,\pi^*,s}(\mathtt{Gain}) = \mathtt{val}^{\mathcal{G}}_{\mathtt{Gain}}(s) \leq \mathtt{val}^{\mathcal{G}}_{\mathtt{EN}(i) \cap \mathtt{EPAR}}(s) + \varepsilon \tag{3.5}$$

For rewards in unary, N is doubly exponential, i.e.,
$N \in \mathcal{O}\big(\exp\big(\exp\big(\|\mathcal{G}\|^{\mathcal{O}(1)}\big)\big) \cdot \log(1/\varepsilon)\big)$ *and it can be computed in exponential time. For rewards in binary, the size of N and its computation time increase by one exponential, respectively.*

Proof. Assume that rewards are in unary. The first inequality of (3.4) holds because $\texttt{EN}(i) \cap \texttt{EPAR} \subseteq \texttt{Gain}$ for any i. The third inequality of (3.4) follows from the definition of the value. Towards the second inequality of (3.4), we consider the minimizing MDP $\mathcal{M}(s) \stackrel{\text{def}}{=} \mathcal{G}^{\sigma^*(s)}$ obtained by fixing the Max strategy $\sigma^*(s)$. Since $\sigma^*(s)$ is optimal for Max from state s w.r.t. the objective $\texttt{Gain}$, Lemma 3.3 yields that

$$\texttt{val}^{\mathcal{G}}_{\texttt{Gain}}(s) = \texttt{val}^{\mathcal{M}(s)}_{\texttt{Gain}}((\mathsf{m}_0, s)). \tag{3.6}$$

Since $\sigma^*(s)$ has $\mathcal{O}(\exp(\|\mathcal{G}\|^{\mathcal{O}(1)}))$ memory modes, the size of $\mathcal{M}(s)$ is exponential in $\|\mathcal{G}\|$ and $\mathcal{M}(s)$ can be computed in exponential time.

Now we consider the dual maximizing MDP $\mathcal{M}(s)^d$ and the objectives $\texttt{Term}(i) \cup \texttt{OPAR}$ and $\texttt{Loss}$. (Note that $\mathcal{M}(s)^d$ has the same size as $\mathcal{M}(s)$.) From Lemma 3.7, we obtain a bound $N(s) \in \mathbb{N}$ such that for all $i \geq N(s)$

$$0 \leq \texttt{val}^{\mathcal{M}(s)^d}_{\texttt{Term}(i) \cup \texttt{OPAR}}((\mathsf{m}_0, s)) - \texttt{val}^{\mathcal{M}(s)^d}_{\texttt{Loss}}((\mathsf{m}_0, s)) \leq \varepsilon. \tag{3.7}$$

By Lemma 3.7 and Lemma 3.6, $N(s)$ is exponential in $\|\mathcal{M}(s)^d\|$ and thus doubly exponential in $\|\mathcal{G}\|$, *i.e.*, $N(s) \in \mathcal{O}\left(\exp(\exp(\|\mathcal{G}\|^{\mathcal{O}(1)})) \cdot \log(1/\varepsilon)\right)$. Moreover, $N(s)$ can be computed in time polynomial in $\|\mathcal{M}(s)^d\|$ and thus in time exponential in $\|\mathcal{G}\|$. By duality, we can rewrite Equation (3.7) for $\mathcal{M}(s)$ as follows. For all $i \geq N(s)$

$$\begin{aligned} 0 \leq \texttt{val}^{\mathcal{M}(s)}_{\texttt{Gain}}((\mathsf{m}_0, s)) - \texttt{val}^{\mathcal{M}(s)}_{\texttt{EN}(i) \cap \texttt{EPAR}}((\mathsf{m}_0, s)) \\ \leq \varepsilon. \end{aligned} \tag{3.8}$$

In order to get a uniform upper bound that holds for all states, let $N \stackrel{\text{def}}{=} \max_{s \in S} N(s)$. Since $|S|$ is linear, we still have $N \in \mathcal{O}\left(\exp(\exp(\|\mathcal{G}\|^{\mathcal{O}(1)})) \cdot \log(1/\varepsilon)\right)$ and it can be computed in exponential time in $\|\mathcal{G}\|$. Finally, we can show the second inequality of (3.4).

$$\begin{aligned} &\inf_{\pi} \mathcal{P}^{\mathcal{G}}_{\sigma^*(s),\pi,s}(\texttt{EN}(i) \cap \texttt{EPAR}) \\ &= \inf_{\pi} \mathcal{P}^{\mathcal{M}(s)}_{\pi,(\mathsf{m}_0,s)}(\texttt{EN}(i) \cap \texttt{EPAR}) \\ &= \texttt{val}^{\mathcal{M}(s)}_{\texttt{EN}(i) \cap \texttt{EPAR}}((\mathsf{m}_0, s)) \\ &\geq \texttt{val}^{\mathcal{M}(s)}_{\texttt{Gain}}((\mathsf{m}_0, s)) - \varepsilon && \text{by } i \geq N \geq N(s) \text{ and Equation (3.8)} \\ &= \texttt{val}^{\mathcal{G}}_{\texttt{Gain}}(s) - \varepsilon && \text{by (3.6)} \end{aligned}$$

The first inequality of (3.5) holds because $\texttt{EN}(i) \cap \texttt{EPAR} \subseteq \texttt{Gain}$ for any i. The equality in (3.5) holds by the optimality of π^*. The second inequality of (3.5) follows from the previously stated consequence of (3.4).

For rewards in binary, the sizes of the numbers $N(s)$ (and hence N) and the time to compute it increase by one exponential by Lemma 3.7. ◀

3.6 Unfolding the Game to Energy Level N

Given an SSG $\mathcal{G} = (S, (S_\Box, S_\Diamond, S_\bigcirc), E, P)$ and error tolerance $\varepsilon > 0$, for each state $s \in S$ and energy level $i \geq 0$, we want to compute a rational number v' which satisfies $0 \leq v' - \mathtt{val}^{\mathcal{G}}_{\mathtt{EN}(i)\cap\mathtt{EPAR}}(s) \leq \varepsilon$, and ε-optimal FDD strategies σ_ε and π_ε for Max and Min, resp. We achieve this by constructing a finite-state parity game $\mathcal{G}'$ that closely approximates the original game $\mathcal{G}$, as described in Section 3.3(Item 3.).

For clarity, we explain the construction in two steps. In the first step, we consider a finite-state parity game $\mathcal{G}[N]$. (Unlike $\mathcal{G}'$, the game $\mathcal{G}[N]$ is not actually constructed. It just serves as a part of the correctness proof.) $\mathcal{G}[N]$ encodes the energy level up-to $N + R$ (where R is the maximal transition reward) into the states, *i.e.*, it has states of the form (s, k) with $k \leq N + R$. It imitates the original game $\mathcal{G}$ until energy level $N + R$, but at any state (s, i) with energy level $i \geq N$ it jumps to a winning state with probability $\mathtt{val}^{\mathcal{G}}_{\mathtt{EN}(i)\cap\mathtt{EPAR}}(s)$ and to a losing state with probability $1 - \mathtt{val}^{\mathcal{G}}_{\mathtt{EN}(i)\cap\mathtt{EPAR}}(s)$. (We need the margin up-to $N + R$, because transitions can have rewards > 1, so the level N might not be hit exactly.) Similarly, at states $(s, 0)$ with energy level 0, we jump to a losing state. The coloring function in the new game $\mathcal{G}[N]$ derives its colors from the colors in the original game $\mathcal{G}$, *i.e.*, all states (s, i) have the same color as s in $\mathcal{G}$.

▶ **Definition 3.9** (Definition of $\mathcal{G}[N]$)**.** *We present formally the definition of the game $\mathcal{G}[N]$, which unfolds the energy level in $\mathcal{G}$ until N*

$$\mathcal{G}[N] \stackrel{def}{=} (S[N], (S_\Box[N], S_\Diamond[N], S_\bigcirc[N]), E[N], P[N])$$

where

1. $S[N] \stackrel{def}{=} S \times \{0, \ldots, N + R\} \uplus \{s_{\text{win}}, s_{\text{lose}}\}$*, the set of states is the tuple with the game state and energy level until $N + R$ as the maximum change in a single step is R and since we are only interested in energy levels $\leq N$, it suffices to consider till $N + R$.*

2. $S_\odot[N] \stackrel{def}{=} S_\odot \times \{1, \ldots, N\}$*, both players control their respective states until energy level N. Every state with energy $> N$ becomes a chance node. Consequently,*

3. $S_{\bigcirc}[N] \stackrel{def}{=} S_{\bigcirc} \times \{1, \ldots, N\} \cup S \times \{0, N+1, \ldots, N+R\} \cup \{s_{\text{win}}, s_{\text{lose}}\}$, *since the Max loses when the energy level becomes* ≤ 0, *we make these states as a chance vertex which go to a losing loop.*

4. $E[N]$, $P[N]$

   (a). *For* $0 < i \leq N$, $(s,i) \longrightarrow (s', \max(0,j))$ *iff* $s \xrightarrow{j-i} s' \in E$, *this is just simulating the transitions of the game until energy level* N *and taking care of border cases. When energy drops below* 0, *we move to level* 0 *as there is no difference. When it shoots above* N, *it cannot go beyond* $N+R$ *and thus the transition is well defined.*

   (b). *If* $s \in S_{\bigcirc}$ *above, then the probability is carried over.*

   (c). $(s,0) \longrightarrow s_{\text{lose}}$ *with probability 1.*

   (d). $(s, N+k) \longrightarrow s_{\text{win}}$ *with probability* $\mathtt{val}^{\mathcal{G}}_{\mathtt{EN}(N+k) \cap \mathtt{EPAR}}(s)$ *and with remaining probability moves to* s_{lose} *for* $1 \leq k \leq R$

   (e). $s_{\text{lose}} \longrightarrow s_{\text{lose}}$ *with probability 1. Similarly for* s_{win}.

By construction of $\mathcal{G}[N]$, for $i \leq N$, the EPAR value of (s,i) in $\mathcal{G}[N]$ coincides with $\mathtt{val}^{\mathcal{G}}_{\mathtt{EN}(i) \cap \mathtt{EPAR}}(s)$.

In the second step, since we do not know the exact values $\mathtt{val}^{\mathcal{G}}_{\mathtt{EN}(i) \cap \mathtt{EPAR}}(s)$ for $N+R \geq i > N$, we approximate these by the slightly larger $\mathtt{val}^{\mathcal{G}}_{\mathtt{Gain}}(s)$. I.e., we modify $\mathcal{G}[N]$ by replacing the probability values $\mathtt{val}^{\mathcal{G}}_{\mathtt{EN}(i) \cap \mathtt{EPAR}}(s)$ for the jumps to the winning state by $\mathtt{val}^{\mathcal{G}}_{\mathtt{Gain}}(s)$. Let $\mathcal{G}'$ be the resulting finite-state parity game. It follows from Lemma 3.8 that $0 \leq \mathtt{val}^{\mathcal{G}}_{\mathtt{Gain}}(s) - \mathtt{val}^{\mathcal{G}}_{\mathtt{EN}(i) \cap \mathtt{EPAR}}(s) \leq \varepsilon$ for $i \geq N$ and $\mathtt{Lval}^{\mathcal{G}}_{\mathtt{EN} \cap \mathtt{EPAR}}(s) = \mathtt{val}^{\mathcal{G}}_{\mathtt{Gain}}(s)$. Thus $\mathcal{G}'$ ε-over-approximates $\mathcal{G}[N]$ and $\mathcal{G}$, and we obtain the following lemma.

▶ **Lemma 3.10.** *For all states* s *and all* $0 \leq i \leq N$

$$\mathtt{val}^{\mathcal{G}[N]}_{\mathtt{EPAR}}((s,i)) = \mathtt{val}^{\mathcal{G}}_{\mathtt{EN}(i) \cap \mathtt{EPAR}}(s), \text{ and}$$
$$0 \leq \mathtt{val}^{\mathcal{G}'}_{\mathtt{EPAR}}((s,i)) - \mathtt{val}^{\mathcal{G}[N]}_{\mathtt{EPAR}}((s,i)) \leq \varepsilon.$$

Now we are ready to prove the main theorem.

▶ **Theorem 3.1.** *Let* $\mathcal{G} = (S, (S_{\Box}, S_{\Diamond}, S_{\bigcirc}), E, P)$ *be an SSG with transition rewards in unary assigned by function* r *and colors assigned to states by function* Col. *For every state* $s \in S$, *initial energy level* $i \geq 0$ *and error margin* $\varepsilon > 0$, *one can compute*

1. *a rational number v' such that $0 \leq v' - \mathtt{val}^{\mathcal{G}}_{\mathtt{EN}(i)\cap\mathtt{EPAR}}(s) \leq \varepsilon$ in 2-*NEXPTIME.
[3]

2. *ε-optimal FDD strategies σ_ε and π_ε for Max and Min, resp., in 2-*NEXPTIME. *These strategies use $\mathcal{O}\big(2\text{-EXP}(\|\mathcal{G}\|) \cdot \log\big(\frac{1}{\varepsilon}\big)\big)$ memory modes.*

For rewards in binary, the bounds above increase by one exponential.

Proof. For $i > N$ we output $v' = \mathtt{val}^{\mathcal{G}}_{\mathtt{Gain}}(s)$, which satisfies the condition by Lemma 3.8. For $i \leq N$ we output $v' = \mathtt{val}^{\mathcal{G}'}_{\mathtt{EPAR}}((s,i))$, which satisfies the condition by Lemma 3.10. By Theorem 3.4, the values $\mathtt{val}^{\mathcal{G}}_{\mathtt{Gain}}(s)$ are rational for all states s. Therefore all probability values in $\mathcal{G}'$ are rational and thus the EPAR values of all states in $\mathcal{G}'$ are rational. Hence our numbers v' are always rational.

By Theorem 3.4, the values $\mathtt{val}^{\mathcal{G}}_{\mathtt{Gain}}(s)$ for all states $s \in S$ can be computed in exponential time. By Lemma 3.8, $N \in \mathcal{O}\left(\exp(\exp(\|\mathcal{G}\|^{\mathcal{O}(1)})) \cdot \log(1/\varepsilon)\right)$ is doubly exponential. Therefore, we can construct $\mathcal{G}'$ in $\mathcal{O}\left(\exp(\exp(\|\mathcal{G}\|^{\mathcal{O}(1)})) \cdot \log(1/\varepsilon)\right)$ time and space. Questions about the parity values of states in $\mathcal{G}'$ can be decided in nondeterministic time polynomial in $\|\mathcal{G}'\|$. Thus the numbers v' are computed in 2-NEXPTIME.

Towards Item 2, we construct ε-optimal FDD strategies σ_ε for Max (resp. π_ε for Min) for $\mathtt{EN}(i) \cap \mathtt{EPAR}$ in $\mathcal{G}$. Let $\hat{\sigma}$ (resp. $\hat{\pi}$) be optimal MD strategies for Max (resp. Min) for the objective EPAR in $\mathcal{G}'$, which exist by Remark 2.1. Since these strategies are MD, they can be guessed in nondeterministic time polynomial in the size $\|\mathcal{G}'\|$, and thus in $\mathcal{O}\left(\exp(\exp(\|\mathcal{G}\|^{\mathcal{O}(1)})) \cdot \log(1/\varepsilon)\right)$ nondeterministic time.

Then σ_ε plays as follows. While the current energy level j (i plus the sum of the rewards so far) stays $< N$, then, at any state s', play like $\hat{\sigma}$ at state (s', j) in $\mathcal{G}'$. Once the energy level reaches a value $\geq N$ at some state s' for the first time, then play like $\sigma^*(s')$ forever. (Recall that $\sigma^*(s')$ is the optimal FDD Max strategy for Gain from state s' from Section 3.5.2.) σ_ε is ε-optimal by Lemma 3.10 and Lemma 3.8. It needs to remember the energy level up-to N while simulating $\hat{\sigma}$. Moreover, $\sigma^*(s')$ needs $\mathcal{O}(\exp(\|\mathcal{G}\|^{\mathcal{O}(1)}))$ memory modes by Theorem 3.4. Finally, it needs to remember the switch from $\hat{\sigma}$ to $\sigma^*(s')$. Since $N \in \mathcal{O}\left(\exp(\exp(\|\mathcal{G}\|^{\mathcal{O}(1)})) \cdot \log(1/\varepsilon)\right)$ dominates the rest, σ_ε uses $\mathcal{O}\left(\exp(\exp(\|\mathcal{G}\|^{\mathcal{O}(1)})) \cdot \log(1/\varepsilon)\right)$ memory modes.

[3] We write "computing a number v' in 2-NEXPTIME" as a shorthand for the property that questions like $v' \leq c$ for constants c are decidable in 2-NEXPTIME.

Similarly, π_ε plays as follows. While the current energy level j stays $< N$, at any state s', play like $\hat{\pi}$ at state (s', j) in $\mathcal{G}'$. Once the energy level reaches a value $\geq N$ (at any state) for the first time, then play like π^* forever (where π^* is the uniform optimal MD Min strategy for $\texttt{Gain}$ from Section 3.5.2.) π_ε is ε-optimal by Lemma 3.10 and Lemma 3.8. While π^* is MD and does not use any memory, π_ε still needs to remember the energy level up-to N while simulating $\hat{\pi}$, and thus it uses $\mathcal{O}\left(\exp(\exp(\|\mathcal{G}\|^{\mathcal{O}(1)})) \cdot \log(1/\varepsilon)\right)$ memory modes.

For rewards in binary, all bounds increase by one exponential via an encoding of $\mathcal{G}$ into an exponentially larger but equivalent game with rewards in unary. ◀

No nontrivial lower bounds are known on the computational complexity of approximating $\texttt{val}^{\mathcal{G}}_{\texttt{EN}(i)\cap\texttt{EPAR}}(s)$. However, even without the parity part, the problem appears to be hard. The best known algorithm for approximating the value of the energy objective (resp. the dual termination objective) runs in NEXPTIME for SSGs with rewards in unary [BBEK13].

As for lower bounds on the strategy complexity, ε-optimal Max strategies need at least an exponential number of memory modes (for any $0 < \varepsilon < 1$) even in maximizing MDPs. This can easily be shown by extending the example in Lemma 3.2(Item 2.) and [GOP11, Fig. 1 and Prop. 4] that shows the lower bound for the $\texttt{Gain}$ objective. First loop in a state with an unfavorable color to accumulate a sufficiently large reward (depending on ε) and then switch to the MDP in [GOP11, Fig. 1 and Prop. 4] to play for $\texttt{Gain}$ (since $\texttt{EN}(i) \cap \texttt{EPAR}$ will be very close to $\texttt{Gain}$ then). Even the latter part requires exponentially many memory modes.

Chapter 4

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

4.1 Overview

This chapter is concerned with the strategy complexity of almost surely winning strategies for the energy–meanpayoff objective in maximizing MDPs. We prove that finite-memory strategies suffice for almost surely winning energy– meanpayoff, *i.e.*, the reward on the transitions is multidimensional, and the objective is to satisfy energy on the 1^{st} dimension and *positive* meanpayoff in the remaining dimensions. The strategies we construct are deterministic and require at most exponential memory. We also prove the corresponding lower bound: there is a family of MDPs for which any almost surely winning strategy, even with randomization, needs exponential memory. Since strategies are exponential, the time for any algorithm to synthesize them is also at least EXPTIME in the size of MDP, but for the simpler problem of computing the almost surely winning set of states, we show a pseudo-polynomial upper bound. In Section 4.3 we state the three results and provide a proof sketch for the existence of finite-memory almost surely winning strategies. Each of the next three sections then deals with the proof of one result. Section 4.4 is for proving that finite memory suffices for almost surely winning strategies. The proof is by a contradiction argument which constructs a finite-memory almost surely winning strategy from a state assuming it has an infinite memory strategy but no finite-memory winning strategy. This is the most technical and complex

part of this chapter and proof of this result almost also gives the exponential upper bound result which is the topic of Section 4.5. Section 4.6 discusses a counter example showing that general strategies are strictly more powerful than finite-memory strategies for the non-strict version of energy–meanpayoff objective. Then, in Section 4.7 we show the matching lower bound even for strategies that use randomization. Finally, Section 4.8 discusses the computational complexity for deciding the existence of an almost surely winning strategy from a state and we prove that it is decidable in pseudo-polynomial time.

Contributions. The results in this chapter are based on [DM24] published at ICALP 2024.

4.2 Related Work & Contributions

The existence of almost surely winning strategies for *MeanPayoff-Parity* in MDPs is decidable in polynomial time [CD11a]. These strategies require only finite memory for MeanPayoff > 0 [GOP11], but infinite memory for MeanPayoff ≥ 0 [CD11a].

The existence of almost surely winning strategies for *Energy-Parity* in MDPs is decidable in $\mathsf{NP} \cap \mathsf{coNP}$ and in pseudo-polynomial time [MSTW17]. (The $\mathsf{NP} \cap \mathsf{coNP}$ upper bound holds even for turn-based stochastic games [MSTW21].) Almost surely winning strategies in MDPs require only finite memory in the special case of Energy-Büchi [CD11a], but infinite memory for Energy-co-Büchi and thus for Energy-Parity [MSTW17]. Nevertheless, we follow a similar proof technique of giving an alternating strategy first which uses infinite memory and truncating the strategy to get a finite-memory strategy which is still winning for Energy-MeanPayoff. In some sense, the reason why truncation works for MeanPayoff and not for Parity can be thought of as strictly positive MeanPayoff behaving more like a Büchi objective *i.e.*, you can always choose to ignore it for a little while and you will still be winning. Recall that ε-optimal strategies for Energy-Parity also require only finite (at most doubly exponential) memory, and the value can be effectively approximated in doubly exponential time (even for turn-based stochastic games) Theorem 3.1. However, the strategy structure differs to the one in this chapter.

The *Energy-MeanPayoff* objective is similar to Energy-Parity, but replaces the

Parity part by a strict MeanPayoff objective for a second reward dimension. I.e., one considers an MDP with 2-dimensional transition rewards, where the Energy condition applies to the first dimension and the strict MeanPayoff condition applies to the second dimension. (It can be generalized to higher dimensions d, where the strict MeanPayoff condition applies to all dimensions $2, 3, \ldots, d$.) This might look like a direct generalization of the Energy-Parity objective, since Parity games are reducible to MeanPayoff games [Pur95, Jur98a]. It is important to note that the reductions described in both [Pur95] and [Jur98a] work for both strict and non-strict MeanPayoff objectives although traditionally the reduction is cited for the non-strict case. However, this reduction does not work in the context of these combined objectives when one considers stochastic systems like MDPs; see below.

In [BFRR17], the authors consider the problem of strategy synthesis for beyond worst case (BWC) MeanPayoff. The BWC-MeanPayoff can be stated as looking for a strategy which *surely* achieves a mean-payoff > 0 and on expectation achieves a mean-payoff $> \nu > 0$. The approach and the objective bear some resemblance to the Energy-MeanPayoff. Firstly, since Energy is a safe objective, almost surely satisfying it is the same as satisfying it surely, therefore we are also looking for a strategy which *surely* satisfies Energy and almost surely satisfies strict MeanPayoff. Furthermore, the synthesised strategy in [BFRR17] alternates between two different strategies, each prioritising one part of the problem, similar to $\sigma^*_{\mathtt{alt},Z_b,Z_g,b}$ in Section 4.4. However, the BWC-MeanPayoff asserts both the conditions on the *same* dimension and asks for achieving a certain expected mean-payoff instead of almost surely satisfying a threshold condition on the mean-payoff. It is well known that maximizing the expected mean-payoff and maximizing the probability that the mean-payoff is strictly positive do not necessarily go hand in hand [CKK17]. It is also interesting to observe that while in the BWC-MeanPayoff setting, infinite memory strategies are strictly more powerful than the finite memory ones [BFRR17, Section 4.11], it is not the case for almost surely winning strategies for Energy-MeanPayoff Section 4.4. But the finite-memory strategies, when they exist, only require a pseudo-polynomial number of memory modes for the BWC-MeanPayoff problem. In contrast, exponential memory is required for the strategies in this chapter Item 3..

[CR15] extend the BWC setting to multiple dimensions. They consider MDPs with d-dimensional rewards, where $d = d_1 + d_2$. The objective requires a strictly positive mean-payoff *surely* in the first d_1 dimensions, and an expected mean-

payoff $> \nu$ in the remaining d_2 dimensions. This objective is strictly stronger than Energy-MeanPayoff. E.g., a MeanPayoff of zero in the first dimension may or may not satisfy the Energy objective, but it never satisfies the objective in [CR15].

Non-stochastic Energy-MeanPayoff games have been studied in [BHRR19]. This is the closest to our problem with the only difference being the arena/ model. It is interesting to note that finite memory (exponential) strategies suffice even in this case, although the proof follows by reduction to multidimensional energy games [JLS15]. They also show that in games with only a single player pseudo-polynomial memory suffices. To the best of our knowledge, it is an open question as to whether exponential memory is necessary for the winning strategies.

The objective studied in [BKN16] aims to maximize the expected MeanPayoff (rather than the probability of it being strictly positive) while satisfying the energy constraint. However, unlike in our work, the reward function has a single dimension (*i.e.*, both criteria apply to the same value) and ε-optimal strategies can require infinite memory.

Our contribution. We consider the Energy-MeanPayoff objective in MDPs with d-dimensional rewards. The first dimension needs to satisfy the Energy condition (never drop below 0), while each other dimension needs to have a *strictly* positive MeanPayoff. We show that almost surely winning strategies for Energy-MeanPayoff require only *finite* memory. This is in contrast to the Energy-Parity objective where almost surely winning strategies require infinite memory in general [MSTW17, Page 4] (even for the simpler Energy-co-Büchi objectives). This also shows that Energy-Parity is not reducible to Energy-MeanPayoff in MDPs, unlike the reduction from Parity to MeanPayoff in [Pur95, Jur98a].

Our results do *not* carry over to Energy-MeanPayoff objectives with *non-strict* inequalities where one just requires a MeanPayoff ≥ 0 almost surely. As we demonstrate in Section 4.6, this requires infinite memory even for the case of $d = 2$.

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

of small transition rewards in $\{-1, 0, +1\}$.

Although almost surely winning strategies are 'exponentially large' in this sense, their existence is still decidable in pseudo-polynomial time; cf. Section 4.8.

4.3 The Main Result

▶ **Theorem 4.1.** *Let $\mathcal{M} = (S, S_\Box, S_\bigcirc, E, P)$ be an MDP with d-dimensional rewards on the edges $\boldsymbol{r} : E \to [-R, R]^d$. For the multidimensional Energy-MeanPayoff objective $\mathtt{EN}_1(k) \cap \mathtt{MP}_{[\mathbf{2},\boldsymbol{d}]}(\mathbf{> 0})$ the following properties hold.*

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

In the following three sections we prove items 1.,2.,3. of Theorem 4.1, respectively.

Here we sketch the main idea for the upper bound. Except in a special corner case where the energy fluctuates only in a bounded region, almost-surely winning strategies for Energy-MeanPayoff can be chosen among some particular strategies that alternate between two modes, playing two different memoryless strategies. This alternation keeps the balance between the Energy-part and the MeanPayoff-part of the objective. This is similar to almost-surely winning strategies for the Energy-Parity objective in [MSTW17]. In one mode, one plays a memoryless randomized strategy that almost surely yields a positive mean payoff in all dimensions (in case of Energy-Parity, instead of mean payoff it satisfies Parity almost surely). This is called the *Gain* phase. Whenever the energy level (the cumulative reward in dimension 1) gets dangerously close to zero, one switches to the other mode and plays a different memoryless strategy that focuses exclusively on getting the energy level up again, while temporarily neglecting the other part of the objective (Parity or Mean payoff, respectively). This is called a *Bailout*. Once the energy level is sufficiently high, one switches back to the Gain phase again. The crucial property is that, except in a null set, only finitely many

Bailouts are required, and thus the temporary neglect of the second part of the objective does not matter in the long run. Such a strategy uses infinite memory, because it needs to remember the unbounded energy level. For Energy-Parity (and even Energy-co-Büchi) this cannot be avoided and finite-memory strategies do not work [MSTW17]. However, for Energy-MeanPayoff one can relax the requirements somewhat. Suppose that one records the stored energy only up to a certain bound b, *i.e.*, one forgets about potential excess energy above b. In that case, one might have to do infinitely many Bailouts with high probability, most of which are unnecessary (but one does not know which ones). However, for a sufficiently large bound b, these superfluous Bailouts occur so infrequently that they do not compromise the MeanPayoff-part of the objective. The critical part of the proof is to show this property and an upper bound on b. Once this is established, one obtains a finite-memory strategy, because it suffices to record the energy level only in the range $[0, b]$ (plus one extra bit of memory to record the current phase, Gain or Bailout).

Note that the argument above is different from the one that justifies finite-memory ε-optimal strategies for *Energy-Parity* in Chapter 3. These also record the energy only in a bounded region, but stop doing Bailouts after the upper bound has been visited. I.e., they do too few Bailouts, and thus incur an ε-chance of losing. In contrast, our almost-surely winning strategies for Energy-MeanPayoff rather do too many Bailouts, but sufficiently infrequently such that they don't compromise the objective.

4.4 Proof of Item 1.

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

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

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

Towards a contradiction, assume that not all configurations are winnable with

a finite-memory strategy. I.e., let $\mathtt{Win}_f(s) \stackrel{\text{def}}{=} \{k \mid s \in \mathtt{AS}_f(\mathtt{O}(k))\}$ denote the energy levels from which one can win almost surely with a *finite-memory* strategy from s, and assume that there is a state $s^\dagger$ such that $i_{s^\dagger} \notin \mathtt{Win}_f(s^\dagger)$. We then construct a finite-memory winning strategy from $s^\dagger$ for $\mathtt{O}(i_{s^\dagger})$, leading to a contradiction. Similar to i_s, let f_s denote the minimal k such that $k \in \mathtt{Win}_f(s)$ and ∞ if there is no such k.

▶ **Definition 4.2.** *We construct a new MDP $\mathcal{M}^*$ which abstracts away all the $\mathtt{Win}_f$ configurations. At every state s, the player gets the option to enter a winning sink state if the energy level is sufficiently large to win with finite memory,* i.e.*, if the current energy level is at least f_s. The states of the MDP $\mathcal{M}^*$ will have two copies of each state s of $\mathcal{M}$, namely s and s'. Moreover, we add a new state s_{win}. All states s' are controlled by $\Box$ and every step $s_1 \longrightarrow s$ in the original MDP $\mathcal{M}$ is now mapped to a step $s_1 \longrightarrow s'$ with the same reward (and the same probability if s_1 was a random state). In s', the player has two choices: he can either go to s with reward $\mathbf{0}$ or go to s_{win} with reward $(-f_s, \mathbf{0})$. The latter choice is only available if $f_s < \infty$. s_{win} is a winning sink where $s_{\text{win}} \longrightarrow s_{\text{win}}$ with reward $\mathbf{1}$,* i.e.*, reward $+1$ in all dimensions.*

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

▶ **Lemma 4.3.** *Let $s \in S$ and $k \in \mathbb{N}$, and consider the objective $\mathtt{O}(k)$. There exists an almost surely winning strategy σ^* from s in $\mathcal{M}^*$ if and only if there exists an almost surely winning strategy σ from s in $\mathcal{M}$. Moreover, if σ^* is finite-memory then σ can be chosen as finite-memory, and vice-versa.*

Proof. Towards the 'only if' direction, let σ^* be a strategy from s in $\mathcal{M}^*$ that is almost surely winning for $\mathtt{O}(k)$. We define a strategy σ from s in $\mathcal{M}$ that plays as follows. First σ imitates the moves of σ^* until (if ever) σ^* chooses a move $s'_1 \to s_{\text{win}}$ with non-zero probability at some state s'_1. This is possible, since any finite path in $\mathcal{M}^*$ that does not contain s_{win} can be bijectively mapped to a path in $\mathcal{M}$. The only difference is that paths in $\mathcal{M}^*$ contain extra steps via primed states, which are skipped in the paths in $\mathcal{M}$. Moreover, the transition probabilities at random states coincide in $\mathcal{M}^*$ and $\mathcal{M}$. If σ^* chooses a move $s'_1 \to s_{\text{win}}$ with non-zero probability at some state s'_1 then the current energy level must be $\geq f_{s_1}$, because σ^* satisfies the energy objective almost surely (and thus

even surely). Thus, in $\mathcal{M}$, there exists an almost surely winning finite-memory strategy $\hat{\sigma}$ for $\mathtt{O}(f_{s_1})$ from s_1. In this situation σ continues by playing $\hat{\sigma}$ from s_1. Therefore, σ satisfies the energy objective surely. Moreover, by shift invariance and the properties of $\hat{\sigma}$, it also satisfies the Mean payoff objective almost surely. Thus, σ satisfies $\mathtt{O}(k)$ almost surely. Finally, if σ^* is finite-memory then so is σ, because $\hat{\sigma}$ is also finite-memory.

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

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

▶ **Lemma 4.4.** *For every state s with $f_s > i_s$ and every $\ell \in \mathbb{N}$, there exists a $\delta_\ell > 0$ such that $\mathtt{val}^{\mathcal{M}^*}_{\mathtt{O}(i_s) \cap \mathtt{Infix}_1(\ell)}(s) \leq 1 - \delta_\ell$.*

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

$\mathtt{O}(i_s) \cap \mathtt{Infix}_1(\ell) = \mathtt{EN}_1(i_s) \cap \mathtt{MP}_{[\mathbf{2},d]}(\mathbf{> 0}) \cap \mathtt{Infix}_1(\ell)$ which is the same as $\mathtt{ST}_1(i_s, \ell) \cap \mathtt{MP}_{[\mathbf{2},d]}(\mathbf{> 0})$. Therefore, we have $\mathtt{val}^{\mathcal{M}^*}_s\big(\mathtt{ST}_1(i_s, \ell) \cap \mathtt{MP}_{[\mathbf{2},d]}(\mathbf{> 0})\big) = 1$. Below we prove that this objective has a finite-memory almost-surely winning strategy σ in $\mathcal{M}^*$. Consider a modified MDP $\mathcal{M}^*_1$ that encodes the energy level up to $i_s + \ell$ in the states. A step exceeding the upper energy bound $i_s + \ell$ results in a truncation to $i_s + \ell$, while a step leading to a negative energy leads to a losing sink. There exists a memoryless randomized (MR) strategy σ_1 in $\mathcal{M}^*_1$ from state (s, i_s) that wins $\mathtt{MP}_{[\mathbf{2},d]}(\mathbf{> 0})$ almost surely, by Lemma 4.6. We can then carry σ_1 back to $\mathcal{M}^*$ as a finite-memory strategy σ with $i_s + \ell + 1$ memory modes such that $\mathcal{P}^{\mathcal{M}^*}_{\sigma,s}\big(\mathtt{ST}_1(i_s, \ell) \cap \mathtt{MP}_{[\mathbf{2},d]}(\mathbf{> 0})\big) = 1$. By set inclusion, $\mathcal{P}^{\mathcal{M}^*}_{\sigma,s}(\mathtt{O}(i_s)) = 1$. By Lemma 4.3, there also exists a finite-memory strategy from s in $\mathcal{M}$ that is almost

surely winning for $\mathtt{O}(i_s)$. This implies $f_s = i_s$, a contradiction to our assumption $f_s > i_s$. Hence, we obtain $\delta_\ell \stackrel{\text{def}}{=} 1 - \mathtt{val}^{\mathcal{M}^*}_{\mathtt{O}(i_s) \cap \mathtt{Infix}_1(\ell)}(s) > 0$. ◀

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

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

▶ **Lemma 4.5.** *[BKN16, Lemma 3] Let $\mathcal{M}$ be an MDP. If $s \in \mathtt{AS}(\mathtt{Bailout}(k))$ for some $k \in \mathbb{N}$ then $i_s^{\mathtt{Bailout}} \leq |S| \cdot R$. Moreover, there exists a uniform MD strategy $\sigma^*_{\mathtt{Bailout}}$ which is almost surely winning $\mathtt{Bailout}(k)$ from every state $s \in \mathtt{AS}(\mathtt{Bailout}(k))$.*

Proof. We rely on the existence of "pumping" strategies established in [BKN16]. In [BKN16, Definition 4], a strategy is defined as *pumping* in a configuration if it is safe (energy never drops below zero on any run) and the energy level tends to infinity almost surely.

We first establish that for finite MDPs and any *finite-memory* strategy, the objective $\mathtt{Bailout}(k)$ is equivalent to the pumping property. A finite-memory strategy induces a finite Markov chain. If such a strategy wins $\mathtt{Bailout}(k)$ almost surely, then $\mathtt{MP}_1 > 0$ almost surely. In a finite Markov chain, this implies that every reachable Bottom Strongly Connected Component (BSCC) must have a strictly positive expected mean payoff. By the Strong Law of Large Numbers, the accumulated energy of a random walk with positive drift diverges to $+\infty$ almost surely, satisfying the pumping condition. Conversely, if a finite-memory strategy is pumping, the energy diverges to $+\infty$. In a finite state space, this implies the strategy cannot be trapped in any BSCC with non-positive expected weight (where energy would either recur or diverge to $-\infty$). Thus, it must eventually reach and stay in BSCCs with positive expected yield, satisfying $\mathtt{MP}_1 > 0$.

[BKN16, Lemma 3] proves that for every EMDP, there exists a uniform memoryless strategy $\sigma^*_{\mathtt{Bailout}}$ that is pumping in every pumpable configuration. Since $s \in \mathtt{AS}(\mathtt{Bailout}(k))$, the configuration admits a winning strategy. As established above, this implies the configuration is pumpable. Thus, the uniform memoryless strategy $\sigma^*_{\mathtt{Bailout}}$ exists. Being memoryless (and thus finite-memory), this pumping strategy is almost-surely winning for $\mathtt{Bailout}(k)$.

Regarding the energy bound, [BKN16, Lemma 3] provides a bound of $3 \cdot \|\mathcal{M}\| \cdot R$ based on a reduction to Energy-Büchi games on an expanded state space. However, a tighter bound of $|S| \cdot R$ holds structurally. The memoryless strategy induces a Markov chain. The requirement of *sure* safety implies that no run in the support of the chain can traverse a cycle with negative total weight (otherwise the energy would drop unboundedly). Since all reachable cycles must be non-negative, the maximum energy deficit is bounded by the deficit accumulated along a simple (loop-free) path. The length of a simple path is bounded by $|S|$, so an initial energy $i_s^{\texttt{Bailout}} \leq |S| \cdot R$ suffices. ◀

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

▶ **Lemma 4.6.** *[BBC⁺14, Proposition 5.1] There is a uniform MR strategy $\sigma^*_{\texttt{Gain}}$ which is almost surely winning for* `Gain` *(or any subset of dimensions) from all states $s \in \texttt{AS}(\texttt{Gain})$.*

Proof. We construct the uniform MR strategy $\sigma^*_{\texttt{Gain}}$ by analysing the Maximal End Components (MECs) of the MDP $\mathcal{M}$. Recall that the objective `Gain` requires $\texttt{MP} > \mathbf{0}$ almost surely.

Let $\mathcal{C}$ be the set of all MECs $C = \left(S_C, S_\Box^C, S_\bigcirc^C, E_C, P_C\right)$ of $\mathcal{M}$. For a MEC C, let $\mathcal{P}_C$ denote the set of all vector values achievable as the expected mean payoff of a strategy within C. As established in [BBC⁺14, Theorem 4.1], the set of achievable expectations is characterized by the system of linear inequalities L. Thus, for a specific MEC, $\mathcal{P}_C$ is the set of vectors satisfying the flow constraints of C (specifically equations 4.3–4.5 restricted to C), which forms a convex polytope.

We define a MEC C to be *winning* if this set $\mathcal{P}_C$ contains a vector $\boldsymbol{w}$ strictly greater than $\mathbf{0}$. If such a $\boldsymbol{w}$ exists, let $\delta = \min_i w_i > 0$. We rely on the construction in the proof of Proposition 5.1 in [BBC⁺14]. The authors show that if a vector $\boldsymbol{v} \in \mathcal{P}_C$ satisfies the flow constraints, then for any $\varepsilon > 0$, there exists a Memoryless Randomized (MR) strategy σ_ε such that the long-run average reward is at least $\boldsymbol{v} - \boldsymbol{\varepsilon}$ almost surely. Applying this with precision $\varepsilon < \delta$ guarantees a uniform MR strategy σ_C on C such that:

$$\mathcal{P}_{\sigma_C, s}(\texttt{MP} \geq \boldsymbol{w} - \boldsymbol{\varepsilon}) = 1 \quad \text{for all } s \in S_C.$$

Since $\varepsilon < \delta$, we have $\boldsymbol{w} - \boldsymbol{\varepsilon} > \mathbf{0}$, and thus $\mathcal{P}_{\sigma_C, s}(\texttt{MP} > \mathbf{0}) = 1$.

Let $U_{\mathsf{win}} = \bigcup\{S_C \mid C \text{ is a winning MEC}\}$. We now prove that a state s belongs to $\mathtt{AS}(\mathtt{Gain})$ if and only if it can reach U_{win} with probability 1.

- **Direction ($\Leftarrow$):** If s can reach U_{win} almost surely, the player can use a memoryless deterministic attractor strategy to reach U_{win}. Upon entering a state in a winning MEC C, the player switches to the MR strategy σ_C. Since MECs are closed under σ_C, the run stays in C and satisfies $\mathtt{Gain}$ almost surely.

- **Direction ($\Rightarrow$):** Suppose $s \in \mathtt{AS}(\mathtt{Gain})$. Any strategy in a finite MDP eventually stabilizes in some End Component (and thus some MEC) with probability 1. If the play stabilizes in a MEC C that is *not* winning, then by definition $\mathcal{P}_C \cap \{\boldsymbol{v} \mid \boldsymbol{v} > \mathbf{0}\} = \emptyset$. Since $\mathcal{P}_C$ is closed (it is a polytope) and disjoint from the open set $(\mathbf{0}, \boldsymbol{\infty})$, it is impossible for the limit-average reward to be strictly positive almost surely within C. Therefore, any almost-sure winning strategy must avoid stabilizing in non-winning MECs, implying it must reach U_{win} almost surely.

Consequently, we can define the global uniform MR strategy $\sigma^*_{\mathtt{Gain}}$ as follows: Let $\sigma_{\mathtt{reach}}$ be a uniform Memoryless Deterministic (MD) attractor strategy on $\mathtt{AS}(\mathtt{Gain}) \setminus U_{\mathsf{win}}$ that reaches U_{win} almost surely.

$$\sigma^*_{\mathtt{Gain}}(s) = \begin{cases} \sigma_C(s) & \text{if } s \in S_C \text{ for some winning MEC } C, \\ \sigma_{\mathtt{reach}}(s) & \text{if } s \in \mathtt{AS}(\mathtt{Gain}) \setminus U_{\mathsf{win}}, \\ \text{arbitrary} & \text{otherwise.} \end{cases}$$

A run starting from $s \in \mathtt{AS}(\mathtt{Gain})$ consistent with $\sigma^*_{\mathtt{Gain}}$ enters a winning MEC C almost surely and subsequently follows σ_C. By the construction derived from [BBC+14], the limit-average reward converges almost surely to a value strictly greater than $\mathbf{0}$. ◀

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

▶ **Lemma 4.7.** *In $\mathcal{M}^*$, there are two uniform memoryless strategies $\sigma^*_{\texttt{Bailout}}$ and $\sigma^*_{\texttt{Gain}}$ which, starting from any state $s \in \texttt{AS}(\texttt{O}(k))$, almost surely satisfy $\texttt{Bailout}(k)$ and $\texttt{Gain}$, respectively.*

Proof. Let $s \in \texttt{AS}(\texttt{O}(k))$. We show that $s \in \texttt{AS}(\texttt{Bailout}(k))$ and $s \in \texttt{AS}(\texttt{Gain})$. The existence of the memoryless strategies $\sigma^*_{\texttt{Bailout}}$ and $\sigma^*_{\texttt{Gain}}$ then follows from Lemma 4.5 and Lemma 4.6, respectively.

We assumed that all states s in $\mathcal{M}$ admit an almost surely winning strategy for Energy-MeanPayoff. By Lemma 4.3, this also holds for all states q in $\mathcal{M}^*$. Let $\sigma^\sharp_q$ denote an almost surely winning strategy from q for $\texttt{O}(i_q)$ in $\mathcal{M}^*$ (without restrictions on memory).

Recall from Section 2.1 that the random variable X_t denotes the state at time t, and Y_t denotes the (d-dimensional) sum of the rewards until time t.

▷ Claim 4.8. For every state $q \in \mathcal{M}^*$ there exists some number of steps $n_q \in \mathbb{N}$ and a probability $p_q > 0$ such that

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

Proof. Towards a contradiction, assume that for all m

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

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

Since $\sigma^\sharp_q$ satisfies $\texttt{EN}_1(i_q)$ almost surely, all runs always satisfy $(Y_j)_1 \geq i_{X_j} - i_q$ for all j. On the other hand, our assumption yields $\mathcal{P}^{\mathcal{M}^*}_{\sigma^\sharp_q,q}\left(\bigcup_{j=0}^{m}(Y_j)_1 > i_{X_j} - i_q\right) = 0$. This implies that $(Y_j)_1 = i_{X_j} - i_q$ for all j. Hence, in all runs the energy fluctuates by at most $\ell \stackrel{\text{def}}{=} 2\max_q i_q$. Thus, $\mathcal{P}^{\mathcal{M}^*}_{\sigma^\sharp_q,q}(\texttt{O}(i_q) \cap \texttt{Infix}_1(\ell)) = 1$. Then Lemma 4.4 implies that $f_q = i_q$. Since $X_0 = q$ we have $f_{X_0} = f_q$ and thus $(Y_0)_1 \geq f_{X_0} - i_q = 0$. This contradicts our assumption, since the second part of the union is surely satisfied. ◁

For any state q, let n_q, p_q denote the values from Claim 4.8.

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

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

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

$\sigma_{\mathtt{Bailout}}$ clearly satisfies $\mathtt{EN}_1(k)$ as it is a mix of energy safe strategies $\left(\sigma_q^\sharp\right)_{q \in S^*}$ and since we are starting from a safe energy level. By Claim 4.8, there is a positive probability (lower-bounded by $\min_q p_q > 0$) that either Item 1. or Item 2. happens in each phase.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The following lemma shows the converse of Lemma 4.7. In $\mathcal{M}^*$, it is always possible to win $\mathtt{O}(i_s)$ almost surely from s by playing a particular strategy $\sigma^*_{\mathtt{alt},Z_b,Z_g}$ which combines the two uniform memoryless strategies $\sigma^*_{\mathtt{Bailout}}$ and $\sigma^*_{\mathtt{Gain}}$. Let Z_b denote the minimal universally safe energy level for `Bailout`, *i.e.*, $Z_b \stackrel{\text{def}}{=} \max_s \min\{k \mid s \in \mathtt{AS}(\mathtt{Bailout}(k))\}$. Moreover, let $Z_g > Z_b$ be a larger energy level at which our strategy switches from $\sigma^*_{\mathtt{Bailout}}$ to $\sigma^*_{\mathtt{Gain}}$.

Similarly to [MSTW17], we define an infinite-memory strategy $\sigma^*_{\mathtt{alt},Z_b,Z_g}$ that always records the current energy level and operates by switching between two phases. It starts by playing $\sigma^*_{\mathtt{Gain}}$ (Gain-phase) if our starting energy level is sufficiently high ($\geq Z_b + R$), and otherwise starts by playing $\sigma^*_{\mathtt{Bailout}}$ (Bailout-phase). In the `Bailout`-phase, the primary goal is to pump the energy level up until it is $\geq Z_g$, and then it switches to the `Gain`-phase. It enters the `Bailout`-phase again if the energy level drops below $Z_b + R$ (in which case it will still be $\geq Z_b$).

▶ **Lemma 4.9.** *There exists a $Z_g \in \mathbb{N}$ such that for every s in $\mathcal{M}^*$ the strategy $\sigma^*_{\mathtt{alt},Z_b,Z_g}$ is almost surely winning for $\mathtt{O}(i_s)$ from s.*

Proof. The parameter Z_g is chosen sufficiently large such that there is a fixed non-zero probability that after every Bailout-phase one never needs another `Bailout`. (Thus, except in a null set there are only finitely many Bailouts.) The existence of such a finite Z_g is guaranteed by the fact that $\lim_{k\to\infty} \mathcal{P}_{\sigma^*_{\mathtt{Gain}},s}(\mathtt{O}(k)) = 1$. (Lemma 2.11). Eventually, except in a null set, $\sigma^*_{\mathtt{alt},Z_b,Z_g}$ plays `Gain` forever, thus satisfying $\mathtt{O}(i_s)$ almost surely from s. ◀

Some combined objectives like Energy-Parity really require infinite memory for almost surely winning strategies [MSTW17]. However, we show that a sufficiently large *finite* memory is enough to win Energy-MeanPayoff almost surely. The idea is to modify the strategy $\sigma^*_{\mathtt{alt},Z_b,Z_g}$ such that it remembers the current energy only in the interval $[0, b]$, for some sufficiently large $b > Z_g$, and ignores any possible excess energy above b. This modified strategy is denoted by $\sigma^*_{\mathtt{alt},Z_b,Z_g,b}$, and it has a finite set of memory modes $[0, b] \times \{0, 1\}$. The $\{0, 1\}$ part is used to remember the current phase ($\mathtt{Gain} = 0$ or $\mathtt{Bailout} = 1$). Then $\sigma^*_{\mathtt{alt},Z_b,Z_g,b}[(u, x)]$ denotes the strategy $\sigma^*_{\mathtt{alt},Z_b,Z_g,b}$ with current memory mode $(u, x) \in [0, b] \times \{0, 1\}$.

The finite bound b on the remembered energy has the effect that $\sigma^*_{\mathtt{alt},Z_b,Z_g,b}$ can no longer guarantee a fixed positive probability of not needing another Bailout after each Bailout-phase. Thus, one might have infinitely many Bailouts with

positive probability. (Most of these are unnecessary, but one cannot be sure which ones). Unlike for Energy-Parity, where using infinitely many `Bailout` phases can compromise the objective, the nature of the $\mathtt{MP}_{[\mathbf{2},\boldsymbol{d}]}(\mathbf{> 0})$ objective allows us to use infinitely many Bailouts with non-zero probability, provided that they happen sufficiently infrequently.

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

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

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

▶ **Lemma 4.10.** *If there are constants $v_i^1 > 0$ and v_i^2 such that, for all $i \in [2,d]$*

and states q

$$\mathcal{E}^{\mathcal{M}^*}_{\sigma^*_{\mathtt{alt},Z_b,Z_g,b}[(Z_g,0)],q}((Y_{T^{\mathtt{Gain}}})_i) \geq v_i^1$$
$$\mathcal{E}^{\mathcal{M}^*}_{\sigma^*_{\mathtt{alt},Z_b,Z_g,b}[(Z_b,1)],q}((Y_{T^{\mathtt{Bailout}}})_i) \geq v_i^2$$
$$v_i^1 + v_i^2 > 0$$

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

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

The following technical Lemma 4.11 shows that the constants v_i^1, v_i^2 from Lemma 4.10 exist. Recall that the finite Markov chains $\mathcal{A}^{\mathtt{Gain}}$ and $\mathcal{A}^{\mathtt{Bailout}}$ are obtained by fixing the memoryless strategies $\sigma^*_{\mathtt{Gain}}$ and $\sigma^*_{\mathtt{Bailout}}$ in $\mathcal{M}^*$, respectively. Let $x_{\min,1}$ and $x_{\min,2}$ denote the minimal occurring non-zero probabilities in these two Markov chains, respectively. (They come from solutions of linear programs and can be chosen as only exponentially small, *i.e.*, described by a polynomial number of bits). The proof works by applying general results about expected first passage times in truncated Markov chains to the induced Markov chains $\mathcal{A}^{\mathtt{Gain}}$ and $\mathcal{A}^{\mathtt{Bailout}}$. The general idea is that in the Gain-phase one has a general up drift in all dimensions, and in particular in the first (energy) dimension. It is thus unlikely to go down very far in the energy dimension, even if the energy is truncated at b. Thus, for a sufficiently large truncation point b (actually $b = Z_g + 1$ suffices), the expected time spent in the Gain-phase is very large relative to the expected time spent in the Bailout phase. More exactly, the former increases exponentially in b, while the latter is polynomial in b. For a sufficiently large b (exponential in $\|\mathcal{M}^*\|$), the condition $v_i^1 + v_i^2 > 0$ is met.

▶ **Lemma 4.11.** *Let $\mu_i > 0$ denote the lower bound on the mean payoff in dimension i in any BSCC in the Markov chain $\mathcal{A}^{\texttt{Gain}}$ with corresponding computable constants c_i, $g_{\texttt{Gain}}$, $h_{\texttt{Gain}}$ ((2.5), (2.9), (2.1)), and let μ denote the lower bound on the mean payoff in the 1^{st} dimension in any BSCC of $\mathcal{A}^{\texttt{Bailout}}$ with the corresponding constants $g_{\texttt{Bailout}}, h_{\texttt{Bailout}}$. All the above constants, except c_i, can be chosen as at most exponential in $\|\mathcal{M}^*\|$ and $1/(1-c_i) \in \mathcal{O}\big(\exp\big(\exp\big(\|\mathcal{M}^*\|^{\mathcal{O}(1)}\big)\big)\big)$.*

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

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

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

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

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

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

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

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

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

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

Figure 4.1: LP for an MEC C for `Gain` [BBC+14, Figure 3]

where $LP_{\mathtt{Gain}}$ (cf. Figure 4.1) is some fixed polynomial in n, w, d and $\log R$. For succinctness, we define $w_{\mathtt{Gain}} \stackrel{\text{def}}{=} LP_{\mathtt{Gain}}(n, w, d, R)$.

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

From Lemmas 4.10 and 2.17, for

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

it suffices to choose k and δ such that

$$
\begin{aligned}
&\left((k+1)\cdot\left(\frac{1}{\delta}-1\right) + \lceil \log_{c_i}(\delta(1-c_i))\rceil\right)\cdot \mu_i \cdot \big(1 - 2g_{\mathtt{Gain}}^k\big) - \\
&\qquad h_{\mathtt{Gain}} - R\cdot\left(\left(n + \frac{2}{1-g_{\mathtt{Gain}}}\right) - \frac{2g_{\mathtt{Gain}}}{(1-g_{\mathtt{Gain}})^2}\right) > \\
&\qquad R\cdot\left(n + \frac{2}{1-g_{\mathtt{Bailout}}} + \frac{Z + h_{\mathtt{Bailout}} + R}{\mu}\right) \quad \forall\, 2 \leq i \leq d
\end{aligned} \tag{4.1}
$$

$$k > n \tag{4.2}$$

$$
\begin{aligned}
&\left((k+1)\cdot\left(\frac{1}{\delta}-1\right) + \lceil \log_{c_i}(\delta(1-c_i))\rceil\right)\cdot \mu_i \geq h_{\mathtt{Gain}} \\
&\qquad\qquad \forall\, 2 \leq i \leq d
\end{aligned} \tag{4.3}
$$

The last set of equations become redundant due to the first one. The left-hand side of (4.1) is the over precision of constant v_1^i and similarly the right-hand side

is that of $-v_i^2$. It is simple to notice that once k is fixed, then v_1^i varies with δ as $C_1 \cdot \frac{1}{\delta} - C_2 \log\left(\frac{1}{\delta}\right) - C_3$ and v_2^i as $-C_4 \log\left(\frac{1}{\delta}\right) - C_5$ for some appropriate constants C_i. To further simplify, W.l.o.g, assume δ is sufficiently small such that

$$R\lceil \log_{c_i}(\delta(1-c_i))\rceil + R \geq h_{\texttt{Gain}}.$$

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

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

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

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

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

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

from (2.1)

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

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

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

Figure 4.2: LP_1: Linear program for steady state probabilities in a BSCC

From (2.9)

$$
\begin{aligned}
1 - g_{\mathtt{Gain}} &= 1 - \exp\left(\frac{-x^n_{\min,1}}{n}\right) \\
&\geq \frac{x^n_{\min,1}}{2n} \left(\text{ Since } 1 - e^{-x} \geq \frac{e-1}{e} \cdot x \geq \frac{x}{2} \text{ for } x \in [0,1] \right)
\end{aligned} \tag{4.4}
$$

$$
\implies R \cdot \frac{2}{1 - g_{\mathtt{Gain}}} \leq \frac{4 \cdot n \cdot R}{x^n_{\min,1}} \tag{4.5}
$$

$$
= 2h_{\mathtt{Gain}} \tag{4.6}
$$

$$
\leq 2\, 2^{f(n,w_{\mathtt{Gain}},d,R)} \tag{4.7}
$$

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

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

To get a lower bound on μ, let $\mathcal{B}$ be any BSCC of $\mathcal{A}^{\mathtt{Bailout}}$ and $P_{\mathcal{B}}$ be the one-step transition probability matrix in $\mathcal{A}^{\mathtt{Bailout}}$ restricted to $\mathcal{B}$. Clearly the number of states in $\mathcal{B}$ is $\leq n$. The steady state probabilities $\pi_{\mathcal{B}}$ are solution to the linear system Figure 4.2.

We apply [Goe94, Theorem 15]. First, lets multiply each row by lcm of denominators to get integer entries. The size of each entry is now bounded by $n \cdot w$. Therefore, size of the entire matrix is $\leq n^3 \cdot w$. $size(b)$ here $\leq 2n + 2$. $\implies$ the denominator of each component of $\pi_{\mathcal{B}}$ is $\leq 2^{\left(n^3 w + 2n + 2\right)} \leq 2^{f^2(n,w,d,R)}$.

The mean payoff in this BSCC is then given by

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

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

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

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

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

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

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

$$\Big((k+1) \cdot \Big(\frac{1}{\delta} - 1 \Big) + \lceil \log_{c_i}(\delta(1-c_i)) \rceil \Big) \cdot \mu_i \cdot \big(1 - 2g^k_{\texttt{Gain}} \big) \geq (k+1) \cdot \Big(\frac{1}{\delta} - 1 \Big) \cdot \frac{\mu_i}{2} \quad (4.8)$$

$$\begin{aligned}
&\geq \frac{k\mu_i}{2\delta} \quad \text{Assume } \delta < \frac{1}{k+1} \\
&\geq 2^{-2f^2(n,w_{\texttt{Gain}},d,R)} \cdot \frac{1}{\delta} \\
&\text{since } k \geq 2 \text{ and } \mu_i \geq 2^{-2f^2(n,w_{\texttt{Gain}},d,R)}
\end{aligned}$$

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

From (2.5) and (2.4)

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

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

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

Combining everything, we have LHS

$$\geq 2^{-2f^2(n,w_{\mathtt{Gain}},d,R)} \cdot \frac{1}{\delta} - 9 \cdot 2^{10f^2(n,w_{\mathtt{Gain}},d,R)} \log_2\left(\frac{1}{\delta}\right) - 99 2^{11f^2(n,w_{\mathtt{Gain}},d,R)} - 2^{3f(n,w_{\mathtt{Gain}},d,R)} \tag{4.9}$$

Comparing it with the constants from Lemma 4.11, one can see that $C_1 = 2^{-2f^2(n,w_{\mathtt{Gain}},d,R)}$, $C_2 + C_3 = 9 \cdot 2^{10f^2(n,w_{\mathtt{Gain}},d,R)}$ and $C_4 + C_5 = 102\, 2^{11f^2(n,w_{\mathtt{Gain}},d,R)}$ all of which are in the required complexity.

Finally, it suffices to choose a δ such that

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

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

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

Proof of Theorem 4.1(Item 1.) Towards a contradiction, we assume that there exists a state $s^\dagger$ such that there is no finite-memory almost surely winning strategy from $s^\dagger$ for $\mathtt{O}(i_{s^\dagger})$ in the MDP $\mathcal{M}$.

First we consider the MDP $\mathcal{M}^*$. The finite-memory strategy $\sigma^*_{\mathtt{alt},Z_b,Z_g,b}[(i_{s^\dagger}, 1)]$ from $s^\dagger$ is energy-safe by construction and satisfies $\mathtt{EN}_1(i_{s^\dagger})$ surely. Now consider the finite Markov chain induced by fixing this finite-memory strategy in $\mathcal{M}^*$. By

Lemma 4.10 and Lemma 4.11, for a sufficiently large (exponential) b it yields a strictly positive expected mean payoff $v_i^1 + v_i^2 > 0$ in every dimension $i \in [2, d]$ in every BSCC of this Markov chain. Since the Markov chain is finite, this implies that the mean payoff in every dimension $i \in [2, d]$ is strictly positive almost surely. Hence, $\mathcal{P}^{\mathcal{M}^*}_{\sigma^*_{\mathtt{alt},Z_b,Z_g,b}[(i_{s^\dagger},1)],s^\dagger}\big(\mathtt{MP}_{[2,d]}(>0)\big) = 1$ and thus $\mathcal{P}^{\mathcal{M}^*}_{\sigma^*_{\mathtt{alt},Z_b,Z_g,b}[(i_{s^\dagger},1)],s^\dagger}(\mathtt{O}(i_{s^\dagger})) = 1$. So there exists an almost surely winning finite-memory strategy from $s^\dagger$ for $\mathtt{O}(i_{s^\dagger})$ in $\mathcal{M}^*$. However, Lemma 4.3 then implies that there also exists an almost surely winning finite-memory strategy from $s^\dagger$ for $\mathtt{O}(i_{s^\dagger})$ in $\mathcal{M}$. Contradiction. ◀

Remark 4.12. If $\sigma^*_{\mathtt{alt},Z_b,Z_g,b}$ satisfies $\mathtt{O}(i_s)$ almost surely from some state s then it also satisfies the stronger objective $\mathtt{O}(i_s) \cap \mathtt{Infix}(b)$ almost surely. Consider a winning run induced by $\sigma^*_{\mathtt{alt},Z_b,Z_g,b}$. While the true energy might sometimes be higher than b, the energy remembered by $\sigma^*_{\mathtt{alt},Z_b,Z_g,b}$ is always $\leq b$. Even with this conservative under-approximation of the energy, the run still satisfies the energy objective. Therefore, in any winning run induced by $\sigma^*_{\mathtt{alt},Z_b,Z_g,b}$, the energy can never *decrease* by more than b. Thus, also $\mathtt{Infix}(b)$ is satisfied almost surely.

4.5 Proof of Item 2.

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

First, inside any BSCC B of $\mathcal{M}^\sigma$, we construct an almost surely winning strategy σ_B and upper bound the minimal safe energy levels and energy fluctuation while following σ_B. Using this, we upper bound the energy fluctuations in paths before reaching a BSCC. We use the fact that the set of states and transitions that occur in any BSCC of a Markov chain induced by fixing some finite-memory strategy in an MDP is an end component of this MDP ([DA97, Theorem 3.2]).

▶ **Lemma 4.13.** *Let B be a BSCC of $\mathcal{M}^\sigma$ and let $\mathcal{M}(B)$ be the corresponding end component in $\mathcal{M}$ with states S_B and transitions E_B. Then there is a strategy σ_B, a bound $b_B \in \mathcal{O}\big(\exp\big(\|\mathcal{M}(B)\|^{\mathcal{O}(1)}\big)\big)$ such that for any state $q \in S_B$, there is a minimal safe energy level $j_q \stackrel{def}{=} i_q^{\mathcal{M}(B)} \leq |S_B| \cdot R$ such that $\mathcal{P}^{\mathcal{M}(B)}_{\sigma_B,q}(\mathtt{O}(j_q) \cap \mathtt{Infix}(b_B)) = 1$.*

Proof Sketch. (Full proof follows) The idea is that for $\mathcal{M}(B)$ there are two cases. In the first case it behaves similar to $\mathcal{M}^*$ from Section 4.4, in the sense that it is possible to win $\texttt{Gain}$ and $\texttt{Bailout}$ almost surely, and thus Energy-MeanPayoff can be won almost surely by switching between the two strategies for $\texttt{Gain}$ and $\texttt{Bailout}$ like in the strategy $\sigma^*_{\texttt{alt},\texttt{Z}_\texttt{b},\texttt{Z}_\texttt{g}}$. Then one can invoke Lemma 4.11 and Remark 4.12 on $\mathcal{M}(B)$ to get an exponential bound b_B such that $\mathcal{P}^{\mathcal{M}(B)}_{\sigma_B,q}(\texttt{O}(j_q) \cap \texttt{Infix}(b_B)) = 1$. If the first case does not hold then $\mathcal{M}(B)$ is very restrictive, and one can show that the energy level fluctuations are bounded by a constant in $\mathcal{O}(|S_B| \cdot R)$.

Before proceeding with the proof of Lemma 4.13, we state some useful definitions and prove some intermediate lemmas which makes it easier to understand the idea. We start by defining the notion of a winning end component (WEC).

▶ **Definition 4.14.** *Let $\mathcal{M}(B) = (S_B, S_{\Box B}, S_{\bigcirc B}, E_B, \boldsymbol{r}_B)$ be an end component of $\mathcal{M}$. We say that $\mathcal{M}(B)$ is a WEC (winning end component) iff there is some strategy $\sigma \in \Sigma_f^{\mathcal{M}(B)}$ such that*

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

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

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

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

▶ **Lemma 4.15.** *If B is a WEC of Type-I, then one can choose σ such that it satisfies all the conditions of Definition 4.14 along with*

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

for every state $q \in S_B$.

Proof of Lemma 4.15. Assume that $\sigma = (\mathsf{M}, \mathsf{m}_0, \mathsf{upd}, \mathsf{nxt})$ which satisfies the requirements of Definition 4.14 gives a mean payoff of 0 in the energy dimension.

Let $\mathfrak{C} = (s_0, \mathsf{m}_0) \xrightarrow{c_0} (s_1, \mathsf{m}_1) \xrightarrow{c_1} \dots (s_k, \mathsf{m}_k) = (s_0, \mathsf{m}_k)$ be a simple cycle of length k in $\mathcal{M}(B)^\sigma$.

▷ Claim 4.16. For any cycle $\mathfrak{C}$ in $\mathcal{M}^\sigma$, $\mathtt{eff}(\mathfrak{C}) = 0$.

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

We construct a strategy σ' which follows Definition 4.14 such that it almost surely satisfies positive meanpayoff in the energy dimension along with $\mathtt{O}(j_s)$ *i.e.*, $\mathcal{P}^{\mathcal{M}(B)}_{\sigma',s}(\mathtt{O}(j_s) \cap \mathtt{MP}_1(> 0)) = 1$. Since B is a WEC of Type-I, there is some cycle $\mathsf{C} = s_0 \xrightarrow{c_0} \dots \xrightarrow{c_{\ell-1}} s_\ell = s_0$ with positive effect. And since effect of every cycle in $\mathcal{M}^\sigma$ is 0, this implies that the reward along *any* path between two given states of $\mathcal{M}^\sigma$ must be identical. Also, every edge in E_B occurs somewhere in $\mathcal{M}^\sigma$ by definition of σ. Let the edge $s_i \xrightarrow{c_i} s_{i+1}$ in C occur at memory mode a_i *i.e.*, $(s_i, \mathsf{a}_i) \xrightarrow{c_i} (s_{i+1}, \mathsf{a}'_i)$. a'_i may or may not be the same as a_{i+1}. However, irreducibility of the Markov chain implies there are paths p_{i+1} connecting (s_{i+1}, a'_i) to $(s_{i+1}, \mathsf{a}_{i+1})$. Consider the cycle $(s_0, \mathsf{a}_0) \xrightarrow{c_0} (s_1, \mathsf{a}'_0) \xrightarrow{\mathtt{eff}(p_1)} (s_1, \mathsf{a}_1) \dots \xrightarrow{\mathtt{eff}(p_\ell)} (s_0, \mathsf{a}_0)$. The effect of the entire cycle is 0 by Claim 4.16 and C is part of this cycle. This implies that the sum of effects of all the paths p_{i+1} is negative and therefore at least one p_{i+1} has negative effect. The strategy σ' simply bypasses this path and updates the memory mode directly to a_{i+1} with some small probability. For some sufficiently small $\varepsilon > 0$

- with probability $(1 - \varepsilon)$ follow σ at (s_i, a_i)
- with probability ε, directly move to $(s_{i+1}, \mathsf{a}_{i+1})$

Observe that it doesn't matter if s_i was a random or a controlled state as the final destination for both edges is the same with only the memory mode being different, so σ' is a valid strategy which updates its memory stochastically.

It is easy to see that σ' also induces an irreducible Markov chain with every edge in E_B occurring at least once, and that the energy objective $\mathtt{EN}(j_s)$ is satisfied as the shortcut introduced has a positive effect on the energy level. Furthermore, for sufficiently small ε, it doesn't change the mean payoff in other dimensions by

much thereby still ensuring that σ' satisfies $\mathtt{MP}_{[2,d]}(> \mathbf{0})$. Finally, the addition of this new edge now causes the mean payoff in 1^{st} dimension to be strictly > 0 as there is now at least one (complex) cycle with positive weight and still no cycles with negative weight from the properties of σ. ◀

Lemma 4.15 shows that it is possible to win both `Gain` and `Bailout` almost surely in $\mathcal{M}(B)$ from every state in $q \in B$ whenever B is a Type-I WEC. I.e., $q \in \mathtt{AS}^{\mathcal{M}(B)}\big(\mathtt{MP}_{[1,d]}(> \mathbf{0})\big)$. Moreover, the minimal safe energy for `Bailout` in $\mathcal{M}(Q)$ from q is exactly j_q, that is $q \in \mathtt{AS}^{\mathcal{M}(B)}(\mathtt{EN}_1(j_q) \cap \mathtt{MP}_1(> 0))$. Thus, $\mathcal{M}(B)$ satisfies the conclusion of Lemma 4.7, *i.e.*, it behaves like $\mathcal{M}^*$.

Therefore, the strategy $\sigma^*_{\mathtt{alt},Z_b,Z_g,b}$, defined in Section 4.4, is almost surely winning $\mathtt{O}(j_q)$ in $\mathcal{M}(B)$. We can now carry over the analysis on the memory bound b for $\mathcal{M}^*$ from Lemmas 4.10 and 4.11 to $\mathcal{M}(B)$. The only difference is that the size is now measured in $\|\mathcal{M}(B)\| \leq \|\mathcal{M}\|$. So we obtain the following lemma.

▶ **Lemma 4.17.** *If B is a WEC of Type-I, there exists a bound $b_B = \mathcal{O}\big(\exp\big(\|\mathcal{M}(Q)\|^{\mathcal{O}(1)}\big)\big)$ such that for all $q \in B$*

$$\mathcal{P}^{\mathcal{M}(B)}_{\sigma^*_{\mathtt{alt},\mathtt{Z_b},\mathtt{Z_g},\mathtt{b_B}},q}(\mathtt{O}(j_q)) = 1.$$

By Remark 4.12, it is also true that $\mathcal{P}^{\mathcal{M}(B)}_{\sigma^*_{\mathtt{alt},\mathtt{Z_b},\mathtt{Z_g},\mathtt{b_B}},q}(\mathtt{O}(j_q) \cap \mathtt{Infix}(b_B)) = 1$.

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

▶ **Lemma 4.18.** *For every WEC B of Type-II, there is a finite-memory strategy σ_B with a constant $b_B \in \mathcal{O}(|S| \cdot R)$ such that*

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

We are now ready to prove Lemma 4.13.

Proof of Lemma 4.13. We provide the bounds based on the type of the end component $\mathcal{M}(B)$. First observe that B is a WEC as σ acts as a witness by satisfying the requirements of Definition 4.14. If B is a WEC of Type-II, then by Lemma 4.18 the minimal energy j_q required to win from any state q in S_B is $\leq |S_B| \cdot R \leq |S| \cdot R$ and the constant b_B is bounded by $\mathcal{O}(|S| \cdot R)$. Choose σ_B and

b_B be the strategy and the constant from Lemma 4.18 in this case. Otherwise, B is a WEC of Type-I. Therefore, j_q in this case would be the same as the minimal energy to satisfy `Bailout` which by Lemma 4.5 is $\leq |S_B|R \leq |S|R$. By choosing σ_B as $\sigma^*_{\mathtt{alt},\mathtt{Z_b},\mathtt{Z_g},\mathtt{b_B}}$, with b_B from Lemma 4.17, we are done. ◀

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

▶ **Lemma 4.19.** *Let T denote the union of all S_B of every BSCC B of $\mathcal{M}^\sigma$, as in Lemma 4.13. Then one can almost surely reach any state in T with the corresponding minimal safe energy level with energy fluctuations of at most $3\cdot|S|\cdot R$.*

Proof of Lemma 4.19. It is clear that σ is also a witness for $\mathtt{EN}_1(i_s) \cap \mathsf{F}\, T$. But by [CD11a, Lemma 2], this can be achieved with at most $2|S|R$ fluctuation in energy. However, since we also need to ensure we maintain the necessary minimal energy level, one can then simply encode the energy level into the state space of $\mathcal{M}$ and enlarge $\mathcal{M}$ up to $(1+2)|S|R$. So the states of this new MDP $\mathcal{M}'$ will now be $S \times [0, 3|S|R]$. Let $T' = \bigcup_{q\in S_B} q \times [i_q^B, 3|S|R]$. Then, it is not hard to see that when starting from (s, i_s) in $\mathcal{M}'$, one almost surely satisfies $\mathsf{F}\, T'$. (Move according to σ until you hit the maximum energy level of $3|S|R$, at which point switch to one of the winning strategies for $\mathtt{EN}_1(\cdot) \cap \mathsf{F}\, T$ which uses only $\mathcal{O}(|S|\cdot R)$ memory modes.) Therefore, one can reach a state q in a BSCC with its safe energy level with a fluctuation of at most $3|S|R$. ◀

Proof of Theorem 4.1(Item 2.) By Lemmas 4.13 and 4.19, for each state s, one can choose a strategy σ and some constant $b \in \mathcal{O}\big(\exp\big(\|\mathcal{M}\|^{\mathcal{O}(1)}\big)\big)$ such that $\mathcal{P}^{\mathcal{M}}_{\sigma,s}(\mathtt{O}(i_s) \cap \mathtt{Infix}(b)) = 1$. This means if one encodes the energy levels between $[0, b]$ into the state space by discarding any excess energy above b and redirecting all the transitions which result in a negative energy to a losing sink (for $\mathtt{MP}_{[2,d]}(>0)$) and constructs this larger MDP $\mathcal{M}[0, b]$, then there is a strategy σ' such that $\mathcal{P}^{\mathcal{M}[0,b]}_{\sigma',(s,k)}\big(\mathtt{MP}_{[2,d]}(>0)\big) = 1$ for every $k \in [i_s, b]$. Then, by Lemma 4.6, there also exists a memoryless (MR) strategy σ^* in $\mathcal{M}[0, b]$ which is almost surely winning $\mathtt{MP}_{[2,d]}(>0)$ from (s, k).

We can carry the memoryless strategy σ^* in $\mathcal{M}[0, b]$ back to $\mathcal{M}$ as a finite-memory strategy $\sigma^*_{\mathcal{M}}$ with memory $[0, b]$. It stores the encoded under-approximated energy level from $\mathcal{M}[0, b]$ in its finite memory instead. Thus, $\sigma^*_{\mathcal{M}}$ is a finite-memory

strategy from s that satisfies $\mathtt{O}(i_s)$ almost surely, and the size of its memory is bounded by $b \in \mathcal{O}\big(\exp\big(\|\mathcal{M}\|^{\mathcal{O}(1)}\big)\big)$.

The strategy $\sigma^*_{\mathcal{M}}$ uses randomization, because σ^* from Lemma 4.6 is MR. However, the MR strategy σ^* for the mean payoff objective could be replaced by a deterministic strategy with an exponential number of memory modes. Hence, the overall number of memory modes in the obtained deterministic version of $\sigma^*_{\mathcal{M}}$ is still only exponential. ◀

4.6 The Boundary of Finite Memory: Non-strict Objectives

The results in Items 1. and 2. establish that finite memory suffices for Energy-MeanPayoff when the MeanPayoff requirement is *strictly* positive (*i.e.*, $\mathtt{MP} > 0$). In this section, we show that this property does not hold for non-strict inequalities. Specifically, for the objective $\mathtt{O}_{\geq} \stackrel{\text{def}}{=} \mathtt{EN}_1(k) \cap \mathtt{MP}_2(\geq 0)$, almost-surely winning strategies may require infinite memory, even in the case of $d = 2$.

Counterexample Construction

We adapt the counterexample for Energy-co-Büchi objectives from [MSTW17, Page 4]. Consider the MDP $\mathcal{M}_{\geq}$ depicted in Figure 4.3. The state space is $S = \{D, A, C, B\}$, where D is a random state and A, C, B are controlled states. The rewards $\boldsymbol{r} \in \mathbb{Z}^2$ on the transitions are defined as follows:

- **From D (Random):**
  - $D \xrightarrow{2/3} A$ with reward $(+1, 0)$.
  - $D \xrightarrow{1/3} C$ with reward $(-1, 0)$.
- **From C (Controlled):**
  - $C \to D$ with reward $(+1, 0)$.
  - $C \to B$ with reward $\mathbf{0}$.
- **From B (Controlled):** $B \to A$ with reward $(0, -1)$.
- **From A (Controlled):** $A \to D$ with reward $\mathbf{0}$.

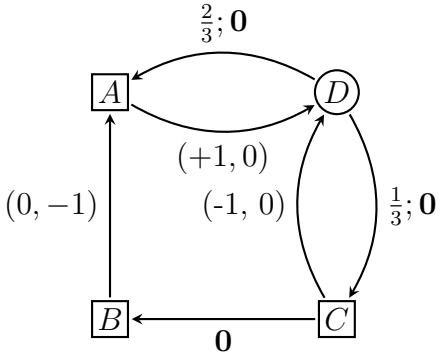

Figure 4.3: The MDP $\mathcal{M}_{\geq}$ inspired by [MSTW17, Page 4]. The reward vectors are $(r_{\texttt{EN}}, r_{\texttt{MP}})$. State B incurs a penalty in the MeanPayoff dimension.

The first dimension represents energy, and the second represents the Mean-Payoff value. The only negative reward in the second dimension is on $B \to A$.

Necessity of Infinite Memory

We prove that winning $\texttt{O}_{\geq}$ almost surely from D (with initial energy $k = 1$) requires infinite memory.

▶ **Proposition 4.20.** *In $\mathcal{M}_{\geq}$, there exists an infinite-memory strategy winning $\texttt{O}_{\geq}$ almost surely from D with initial energy 1, but no finite-memory strategy can do so.*

Proof. Infinite-Memory Strategy. Consider the strategy σ_∞ that tracks the current energy level e. At state C, σ_∞ plays:

$$\sigma_\infty(C) = \begin{cases} B & \text{if } e = 1 \\ D & \text{if } e > 1 \end{cases}$$

At all other states, the choice is deterministic.

Energy Safety: The only transition with negative energy cost is $C \to D$ (cost -1). If the current energy at C is $e > 1$, choosing $C \to D$ results in energy $e - 1 > 0$, which is safe. If $e = 1$, the strategy avoids this edge and chooses $C \to B \to A \to D$. The energy accumulated along $C \to B \to A \to D$ is $0 + 0 + 1 = +1$. Thus, the energy always stays above zero.

MeanPayoff: Whenever $e > 1$, the system behaves as a random walk on the energy levels. From D, the expected change in energy for one return to D is:

$$\mathcal{E}(\Delta e) = \frac{2}{3}(1) + \frac{1}{3}(-1) = \frac{1}{3} > 0.$$

Since the drift is strictly positive, the random walk is transient. Specifically, the probability of the energy level ever returning to 1 (and thus requiring a visit to B) decreases as the energy increases. By the standard properties of transient random walks (Gambler's Ruin on an infinite domain with positive drift), the energy level 1 is visited only finitely often almost surely. Consequently, the state B is visited finitely often almost surely. Since the only negative reward in dimension 2 is on the transition from B, the total accumulated penalty is finite. Thus, the limit average reward is 0, which satisfies $\mathtt{MP}_2(\geq 0)$.

Note that σ_∞ uses infinite memory since it is remembering the exact energy level.

Finite-Memory Failure. Suppose there is a winning finite-memory strategy σ_f. This induces a finite Markov chain $\mathcal{A}$. Consider any Bottom Strongly Connected Component (BSCC) S reachable in $\mathcal{A}$.

- **Case 1: S contains B.** In this case, B is visited infinitely often with a strictly positive limit frequency $\pi_B > 0$. The transition $B \to A$ incurs a reward of -1 in the second dimension, while all other transitions yield 0. The MeanPayoff is $-\pi_B < 0$. Thus, σ_f loses the MeanPayoff objective almost surely.

- **Case 2: S does not contain B.** Then, eventually, the run stays in $\{D, A, C\}$ and always chooses $C \to D$. This induces a random walk on the energy level with steps defined by the transition probabilities at D. Even with positive drift, a random walk on $\mathbb{N}$ starting at any fixed finite energy k has a strictly positive probability of hitting 0 (ruin) if the step sizes are bounded and probabilities are non-trivial. Since σ_f cannot "bail out" to B (as $B \notin S$), the energy objective fails with positive probability.

Since any finite-memory strategy must eventually settle in a BSCC, it must either fail the MeanPayoff objective or the Energy objective. ◀

4.7 The Lower Bound (Proof of Item 3.)

In the previous sections we have shown that finite memory suffices for almost surely winning strategies for the Energy-MeanPayoff objective. However, the required memory depends on the given MDP. We show that no fixed finite amount of memory is sufficient for all MDPs. In fact, the required memory is exponential

in the transition probabilities even for an otherwise fixed 5-state MDP with just one controlled state, $R = 1$ and $d = 2$.

▶ **Definition 4.21.** *Let $1 > \delta > 0$ and $\mathcal{M}_\delta = (S, S_\Box, S_\bigcirc, E, P)$ be an MDP with 2-dimensional rewards. It has just one controlled state s with transitions $s \to s_l$ and $s \to s_r$. From s_l there are two transitions $e_1 = (s_l \to s_l^1)$ and $e_2 = (s_l \to s_l^2)$. Let $\mathcal{P}(e_1) = (1+\delta)/2$ and $\mathcal{P}(e_2) = (1-\delta)/2$ and $\boldsymbol{r}(e_1) = (+1, +1)$ and $\boldsymbol{r}(e_2) = (-1, -1)$. s_l^1 and s_l^2 are random states which each have just one transition back to s with probability 1 and reward $\mathbf{0}$. From s_r there is only one transition e_3 back to s with probability 1 and $\boldsymbol{r}(e_3) = (+1, -1)$.*

The following lemma directly implies the exponential lower bound on the number of memory modes in Theorem 4.1(Item 3.).

▶ **Lemma 4.22.** *Consider the Energy-MeanPayoff objective. For every finite bound $m \in \mathbb{N}$ on the number of memory modes there exists a $\delta \stackrel{def}{=} 1/(6m) > 0$ such that the finite MDP $\mathcal{M}_\delta = (S, S_\Box, S_\bigcirc, E, P)$ from Definition 4.21 satisfies the following properties.*

1. *$\exists \sigma'\ \mathcal{P}^{\mathcal{M}_\delta}_{\sigma',s}(\mathtt{EN}_1(0) \cap \mathtt{MP}_{\mathbf{2}}(> 0)) = 1$,* i.e., *it is possible to win almost surely from s in $\mathcal{M}_\delta$, even with initial energy 0.*

2. *For every finite-memory strategy σ with $\leq m$ memory modes we have $\mathcal{P}^{\mathcal{M}_\delta}_{\sigma,s}(\mathtt{EN}_1(k) \cap \mathtt{MP}_{\mathbf{2}}(> 0)) = 0$ for every $k \in \mathbb{N}$,* i.e., *σ attains nothing in $\mathcal{M}_\delta$, regardless of the initial energy k.*

3. *For $\mathcal{M}_\delta$ we have $|S| = 5$, $d = 1$ and $R = 1$. The number of memory modes required for an almost-surely winning strategy in $\mathcal{M}_\delta$ is exponential in $\|P\|$ (and in $\|\mathcal{M}_\delta\|$).*

Proof. Towards item 1, consider a strategy σ' that plays as follows. It keeps a counter that records the current energy, which is initially 0. Whenever the current energy is 0, it plays $s \to s_r$, otherwise it plays $s \to s_l$. Thus σ' satisfies $\mathtt{EN}_1(0)$ surely from s. Since $\delta > 0$ it follows from the classic Gambler's ruin problem (with strictly positive expected gain, here in the first reward dimension) that σ' plays $s \to s_r$ only finitely often, except in a null set of the runs.Therefore, the expected mean payoff (in the second dimension) under σ' is $(1 + \delta)/2 - (1 - \delta)/2 = \delta > 0$. Hence $\mathcal{P}^{\mathcal{M}_\delta}_{\sigma',s}(\mathtt{MP}_{\mathbf{2}}(> 0)) = 1$. Since the energy objective is satisfied surely, we obtain $\mathcal{P}^{\mathcal{M}_\delta}_{\sigma',s}(\mathtt{EN}_1(0) \cap \mathtt{MP}_{\mathbf{2}}(> 0)) = 1$.

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

Towards item 3, the size of $\mathcal{M}_\delta$ follows from Definition 4.21. By items 1 and 2, the required number of memory modes m for an almost-surely winning strategy satisfies $m > 1/(6\delta)$. Since $\|P\| = \Theta(\log(1/\delta))$ and $\|\mathcal{M}_\delta\| = \Theta(\|P\|)$, we obtain $m = \Omega(\exp(\|P\|))$ and $m = \Omega(\exp(\|\mathcal{M}_\delta\|))$. ◀

The exponential lower bound on the required memory does not require probabilities encoded in binary like in Lemma 4.22. One can construct an equivalent example with polynomially many states where all transition probabilities are $1/2$. This is because one can encode exponentially small probabilities 2^{-k} with a chain of k extra states and transition probabilities $1/2$.

4.8 Computational Complexity

We have shown that the existence of an almost surely winning strategy for the Energy-MeanPayoff objective for a given state and initial energy level in an MDP implies the existence of a deterministic such strategy with exponentially many memory modes (unlike for Energy-Parity which requires infinite memory in general [MSTW17]).

A related problem is the decidability of whether a given state in an MDP and a given initial energy level admit an almost surely winning strategy for Energy-MeanPayoff. This problem is decidable in *pseudo-polynomial* time, using an algorithm very similar to the one for Energy-Parity presented in [MSTW17]. I.e., the time is polynomial, provided that the bound R on the rewards is given in unary. Transition probabilities in the MDP can still be represented in binary.

The crucial point is that it suffices to witness the mere *existence* of an almost surely winning strategy, regardless of its memory. Basically, it suffices that the algorithm proves that the infinite-memory strategy $\sigma^*_{\mathtt{alt},Z_b,Z_g}$ wins almost surely (plus a small extra argument about a corner case where the energy fluctuates only in a bounded region). The algorithm does not need to compute the bound b or to explicitly construct the finite-memory strategy $\sigma^*_{\mathtt{alt},Z_b,Z_g,b}$. The complete decision procedure, combining the bounded case analysis and the greatest fixed point computation for the unbounded case, is detailed in **??** 1.

▶ **Proposition 4.23.** *Let $\mathcal{M} = (S, S_\Box, S_\bigcirc, E, P)$ be an MDP with d-dimensional rewards on the edges $\boldsymbol{r} : E \to [-R, R]^d$. For any state s and $k \in \mathbb{N}$, the existence of an almost surely winning strategy from s for the multidimensional Energy-MeanPayoff objective $\mathtt{EN}_1(k) \cap \mathtt{MP}_{[\mathbf{2},\boldsymbol{d}]}(\mathbf{> 0})$ is decidable in pseudo-polynomial time (*i.e.*, polynomial for R in unary).*

Proof. The proof is similar to the one for Energy-Parity presented in [MSTW17]. The decidability relies on identifying a *stable* winning region $W^\dagger$ where the alternating strategy $\sigma^*_{\mathtt{alt}}$ is viable. This requires that from any state in $W^\dagger$, the player can win both `Gain` and `Bailout` without leaving $W^\dagger$, or safely switch to the bounded winning mode.

Step 1: Bounded Winning Mode (Type-II). We first analyze the case where the objective is won with bounded energy. Any such winning run consists of a transient phase followed by a Recurrent set (Type-II WEC) where energy fluctuations are bounded. By Lemma 4.19 (transient phase) and Lemma 4.18 (recurrent phase), the total energy required is bounded by $B_{\text{tot}} \in \mathcal{O}(|S| \cdot R)$. We compute the function $h : S \to \mathbb{N} \cup \{\infty\}$, where $h(s)$ is the minimal initial energy required to win $\mathtt{MP}_{[\mathbf{2},\boldsymbol{d}]}(\mathbf{> 0})$ almost surely while maintaining energy within $[0, B_{\text{tot}}]$. This is computable in pseudo-polynomial time on the state space $S \times [0, B_{\text{tot}}]$.

Step 2: The Augmented MDP $\mathcal{M}^\dagger$. We construct $\mathcal{M}^\dagger$ from $\mathcal{M}$ by adding a winning sink s_{win}. For every s with $h(s) < \infty$, we add a transition $s \to s_{\text{win}}$ with reward $(-h(s), \mathbf{0})$. This matches the structure of the theoretical $\mathcal{M}^*$ from Lemma 4.3, but replaces the unknown finite-memory thresholds f_s with the computable bounded thresholds $h(s)$. Since any winning strategy is either bounded (captured by h) or pumping (captured by the alternating strategy), this approximation suffices.

Step 3: Greatest Fixed Point Iteration. We identify the stable region

via the greatest fixed point of an operator Ψ. Let $U \subseteq S$. We define the restriction $\mathcal{M}^\dagger[U]$ as the MDP where we retain states $U \cup \{s_{\text{win}}\}$. Crucially, for the almost-sure setting:

- A random vertex $s \in U \cap S_{\bigcirc}$ is retained in $\mathcal{M}^\dagger[U]$ if and only if *all* its transitions in $\mathcal{M}^\dagger$ lead to $U \cup \{s_{\text{win}}\}$. If any transition leaves this set, s is removed (losing).
- A player vertex $s \in U \cap S_{\Box}$ is retained if *at least one* transition leads to $U \cup \{s_{\text{win}}\}$.

We define $\Psi(U) \stackrel{\text{def}}{=} \mathtt{AS}_{\mathcal{M}^\dagger[U]}(\mathtt{Gain}) \cap \mathtt{AS}_{\mathcal{M}^\dagger[U]}(\mathtt{Bailout})$.

- $\mathtt{AS}(\mathtt{Gain})$ is computable in polynomial time (Lemma 4.6).
- $\mathtt{AS}(\mathtt{Bailout})$ is computable in pseudo-polynomial time. Since $\mathtt{Bailout}(k) = \mathtt{EN}_1(k) \cap \mathtt{MP}_1(> 0)$, we also compute the minimal energy $g_U(s)$ required to win $\mathtt{Bailout}$ in $\mathcal{M}^\dagger[U]$. By Lemma 4.5, $g_U(s) \leq |S^\dagger[U]| \cdot R$.

The sequence $W_{i+1} = \Psi(W_i)$ converges to a fixed point $W^\dagger$. Let $g^*(s)$ denote the minimal energy for $\mathtt{Bailout}$ in $W^\dagger$. A state-energy pair (s, k) is winning if and only if $(s \in W^\dagger \wedge k \geq g^*(s)) \vee k \geq h(s)$. ◀

Algorithm 1: Deciding Almost-Sure Winning for Energy-MeanPayoff

Input : MDP $\mathcal{M}$, state s, initial energy k

Output : TRUE if $s \in \mathtt{AS}(\mathtt{EN}_1(k) \cap \mathtt{MP}_{[\mathbf{2},\boldsymbol{d}]}(\mathbf{> 0}))$, else FALSE

`// 1. Compute Minimal Bounded Winning Energy (Type-II)`

$B_{\text{tot}} \leftarrow \text{Poly}(|S|, R)$; `// Bound from Lemma 4.19 + Lemma 4.18`

Construct $\mathcal{M}_{\text{exp}}$ with states $S \times [0, B_{\text{tot}}]$;

Compute winning set W_{exp} for $\mathtt{MP}_{[\mathbf{2},\boldsymbol{d}]}(\mathbf{> 0})$ in $\mathcal{M}_{\text{exp}}$;

foreach $q \in S$ **do**

  $h(q) \leftarrow \min\{e \mid (q, e) \in W_{\text{exp}}\}$ (set to ∞ if empty);

`// 2. Construct Augmented MDP`

$\mathcal{M}^{\dagger} \leftarrow \mathcal{M} \cup \{s_{\text{win}}\}$;

foreach $q \in S$ *such that* $h(q) < \infty$ **do**

  Add transition $q \xrightarrow{(-h(q),\mathbf{0})} s_{\text{win}}$ to $\mathcal{M}^{\dagger}$;

`// 3. Greatest Fixed Point for Unbounded Case`

$W_{\text{curr}} \leftarrow S$; $W_{\text{prev}} \leftarrow \emptyset$;

$g(q) \leftarrow \infty$ for all $q \in S$;

while $W_{curr} \neq W_{prev}$ **do**

  $W_{\text{prev}} \leftarrow W_{\text{curr}}$;

  $\mathcal{M}_{\text{sub}} \leftarrow \text{RestrictAS}(\mathcal{M}^{\dagger}, W_{\text{curr}} \cup \{s_{\text{win}}\})$; `// Remove random states leaking out of` W_{curr}

  $S_{\mathtt{Gain}} \leftarrow \mathtt{AS}(\mathtt{Gain})$ on $\mathcal{M}_{\text{sub}}$;

  $(S_{\mathtt{Bailout}}, g_{\text{new}}) \leftarrow \text{SolveBailout}(\mathcal{M}_{\text{sub}})$; `// Returns winning set and min energy map` g

  $W_{\text{curr}} \leftarrow S_{\mathtt{Gain}} \cap S_{\mathtt{Bailout}}$;

  $g \leftarrow g_{\text{new}}$;

return $(s \in W_{\text{curr}} \wedge k \geq g(s)) \vee k \geq h(s)$;

Chapter 5

Mean-Payoff-Parity and Lifting Strategies from MDPs to 2-Player Stochastic Games

5.1 Overview

In Chapters 3 and 4, we saw methods which are designed w.r.t. a *particular* conjunction of objectives. Fixing an arena model, and a question, it is a priori unclear if there is a single proof technique which works for both energy–parity and energy–meanpayoff simultaneously. For example, if one considers the problem of computing the almost surely winning set of states and the respective strategies in a maximizing MDP for both objectives, the proofs have largely similar structure and ideas [MSTW17, DM24]. Indeed, the latter paper is inspired by the former one. At the same time, both the proofs also consist of elements which are particular to their objectives. Abstracting out these differences could help us unify these proofs and understand the objectives better. Some previous works of this flavor assume some general conditions the objective satisfies and prove results on the strategy complexity for these objectives. We extend some results in this direction in this chapter. Specifically, we consider the strategy complexity *with randomization*. [GK23, MSTW21] lift finite-memory *deterministic update* strategies for Max from MDPs to 2-player stochastic games for shift-invariant inverse-submixing objectives. Their lifting causes an exponential blow up in the number of memory modes required in games. In Section 5.3, we first observe that the lifting technique proposed also generalizes to strategies with stochastic update

functions. We also show a matching lower bound, *i.e.*, the extra exponential memory is required in general. Though these lifting style theorems are usually very useful since it means one only has to deal with the simpler model of MDP, we show the downside of applying such a general theorem. For positive-meanpayoff–parity ($\texttt{MP} > 0 \cap \texttt{EPAR}$) objective which is shift-invariant and inverse-submixing, FDD optimal strategies with exponential memory in maximizing MDPs are shown to exist in [GOP11]. They also give a matching lower bound. By the lifting result, this implies that with deterministic strategies, exponential memory is both necessary and sufficient for Max in stochastic games to play optimally for $\texttt{MP} > 0 \cap \texttt{EPAR}$. We consider what happens when one uses randomization. In MDPs, we show that *memoryless randomized* strategies suffice to play optimally. However, lifting these strategies with improved memory bounds still produces an exponential number of memory modes in games. We show that this is excessive for $\texttt{MP} > 0 \cap \texttt{EPAR}$, optimal strategies only need memory modes at most #(distinct even colors). These strategies use randomization in their memory updates which wasn't necessary in the case of MDPs. We also show a family of games where at least this much of memory is required to play optimally, thereby proving that the bounds are tight. The bit size of the optimal strategy was shown to have exponential dependency on k, the number of even colors. These results are presented in Section 5.4. Finally, in Section 5.5, we consider a different lifting strategy based on [GZ09, BORV23] which lift memoryless (resp. finite-memory) *deterministic* strategies from MDPs (resp. 1-player games) to 2-player games cannot be generalized even to memoryless *randomized* strategies, even under very strong assumptions. I.e., we present stronger counterexamples than the ones given in [BORV23, Van23].

Contributions. The results in this chapter are based on the work submitted to STACS 2026.

5.2 Related Work & Contributions

We consider the *strategy complexity* (*i.e.*, the required memory and randomization) of optimal strategies in SSGs. Given an objective, how does the strategy complexity in SSGs compare to the strategy complexity in MDPs, and can optimal strategies in MDPs be adapted (aka lifted) to work in SSGs.

Lifting strategies from MDPs to SSGs for shift-invariant inverse-submixing

objectives. Optimal and ε-optimal finite-memory (randomized or deterministic) Max strategies for shift-invariant inverse-submixing objectives can be lifted from MDPs to SSGs with an exponential increase in the number of memory modes (using n extra bits of memory in binary-branching games where n states are Min-controlled) [GK23, Theorem 1.2]. (A restricted subcase of this was observed in [MSTW21, Theorem 6].) Note that the assumption of a shift-invariant inverse-submixing Max objective implies that Min has optimal memoryless deterministic (MD) strategies in MDPs/SSGs by [GK23, Theorem 1.1], since this objective is then shift-invariant and submixing for Min.

In Section 5.3 we describe a class of examples for the multi-dimensional $\mathtt{MP} > \mathbf{0}$ objective that shows a corresponding lower bound, *i.e.*, the extra exponential memory for Max in SSGs is required in general. This solves the question in [GK23, end of Sec. 6.1]. Though Max has optimal memoryless randomized (or, alternatively, finite-memory deterministic) strategies in all MDPs, optimal Max strategies in deterministic 2-player games require an exponential number of memory modes, even if randomization is allowed.

Strategy Complexity of Mean-Payoff-Parity. The *mean-payoff-parity* objective is defined as $\mathtt{MP} > 0 \cap \mathtt{EPAR}$. The objective is to attain a strictly positive mean payoff while also satisfying a parity objective. It is shift-invariant inverse-submixing, and hence suitable for the lifting of Max strategies from MDPs to SSGs as in [GK23, Theorem 1.2]. With deterministic strategies, Max already requires an exponential number of memory modes even in MDPs [GOP11, Fig. 1], and hence the extra memory used in the lifting construction (n extra bits of memory for n Min-controlled binary-branching states) still yields an exponential upper bound on the number of memory modes required in SSGs.

The situation is different for randomized strategies. In Section 5.4 we show that Max has optimal *memoryless randomized* strategies in MDPs. However, the lifting construction of [GK23, Theorem 1.2] would still yield exponentially many memory modes for Max strategies in SSGs. We show that optimal Max strategies for $\mathtt{MP} > 0 \cap \mathtt{EPAR}$ in SSGs require (at least and at most) just *polynomially* many memory modes, equal to the number of even colors. I.e., $\mathtt{MP} > 0 \cap \mathtt{EPAR}$ is easier than the exponential worst case for the lifting construction demonstrated in our lower bound. This is an interesting example where randomization in strategies drastically reduces the amount of memory required, but does not eliminate the need for memory entirely.

Comparison with Non-strict Mean-Payoff-Parity. Note however that none of this applies to the *different* mean-payoff-parity objective with a *non-strict* inequality $\texttt{MP} \geq 0 \cap \texttt{EPAR}$. Optimal strategies for this objective are known to require *infinite memory* even in deterministic 1-player games (and thus also in MDPs/SSGs) [CHJ05], due to the possibility of satisfying the non-strict mean-payoff condition by taking the negative rewards infrequently. So, even though the total reward tends to $-\infty$, one can achieve a mean-payoff of 0 by taking longer and longer durations to go down. However, such runs are not winning for *strict* positive mean-payoff which is the main distinction between non-strict and strict versions of mean-payoff.

Lifting deterministic strategies from MDPs to SSGs. A different lifting construction with orthogonal preconditions was described in [GZ09, Theorem 9]. Given an objective, if the players have optimal memoryless *deterministic* strategies in all maximizing (resp. minimizing) MDPs, then they also have optimal memoryless deterministic strategies in all SSGs. Under mild assumptions, this can be generalized to finite-memory deterministic strategies [BORV23, Theorem 4.1] (in stochastic or non-stochastic arenas).

Such results do *not* generalize to *randomized* strategies, even under very strong assumptions. In Section 5.5 we present stronger counterexamples than the ones given in [BORV23, Van23].

5.3 Lifting Strategies from MDPs to 2-Player Stochastic Games for Shift-Invariant Inverse-Submixing Objectives

The *finite memory transfer theorem* of Gimbert & Kelmendi [GK23, Theorem 1.2] shows that finite-memory Min (resp. Max) strategies can be lifted from MDPs to 2-player stochastic games if the payoff function is both shift-invariant and submixing (resp. inverse-submixing). This is a consequence of the following slightly stronger theorem. (Adapted to our notation, because we consider objectives from Max's point of view.)

▶ **Theorem 5.1** ([GK23, Theorem 6.1]). *Let f be a shift-invariant and inverse-submixing payoff function. If, for all $\varepsilon > 0$, Max has an ε-optimal strategy with finite memory in every finite-state MDP, then in every finite-state turn-based*

two-player stochastic game he has an ε-subgame-perfect strategy that has finite memory.

The statement also holds for $\varepsilon = 0$, that is: if Max has a finite-memory optimal strategy in every game controlled by himself, then in every two-player game he has a subgame-perfect strategy with finite memory.

The proof of [GK23, Theorem 6.1] also shows that deterministic (resp. randomized) Max strategies in MDPs are lifted to deterministic (resp. randomized) Max strategies in the game. While [GK23] only consider strategies with deterministic memory updates, their construction can also lift (from MDPs to games) strategies with randomized memory updates. However, in [GK23, Theorem 6.1], the *extra* memory bits used by the constructed Max strategy in the game are always updated deterministically.

The size of the memory used by Max's strategy in the 2-player game is the size of the memory needed in some derived MDPs of a smaller size than the game, plus $k \cdot \lceil \log_2 d \rceil$ extra *bits* of memory, where k is the number of Min-controlled states and d is the maximal out-degree of these states. Thus, in general, Max uses an *exponential* number of memory modes in the game, even if his good strategies in all MDPs are memoryless (or use just polynomially many memory modes); cf. [GK23, Sec. 6.1].

We show a corresponding exponential lower bound in Theorem 5.2. Even if Max has optimal memoryless strategies in all MDPs, in general he still needs exponential memory in the 2-player game, even if randomization is allowed. This solves the question at [GK23, end of Sec. 6.1]. On the other hand, for some objectives the construction in [GK23, Theorem 6.1] uses more memory than necessary. In Section 5.4 we show that for the $\texttt{MP} > 0 \cap \texttt{EPAR}$ objective Max requires, at least and at most, a *polynomial* number of memory modes in stochastic games. The lower bound holds even for the deterministic games.

The multi-dimensional $\texttt{MP} > \mathbf{0}$ objective is shift-invariant (by definition) and inverse-submixing (by [BBE10b, Lemma 6] for dimension 1, and the fact that event-based inverse-submixing objectives are closed under intersection). Moreover, by [BBC+14, Prop. 5.1], almost surely winning strategies for $\texttt{MP} > \mathbf{0}$ in MDPs can be chosen as memoryless randomized. However, the games in Figure 5.1 (similar to [CRR14, Fig. 4] but with tolerance) show that even randomized Max strategies need exponentially many memory modes.

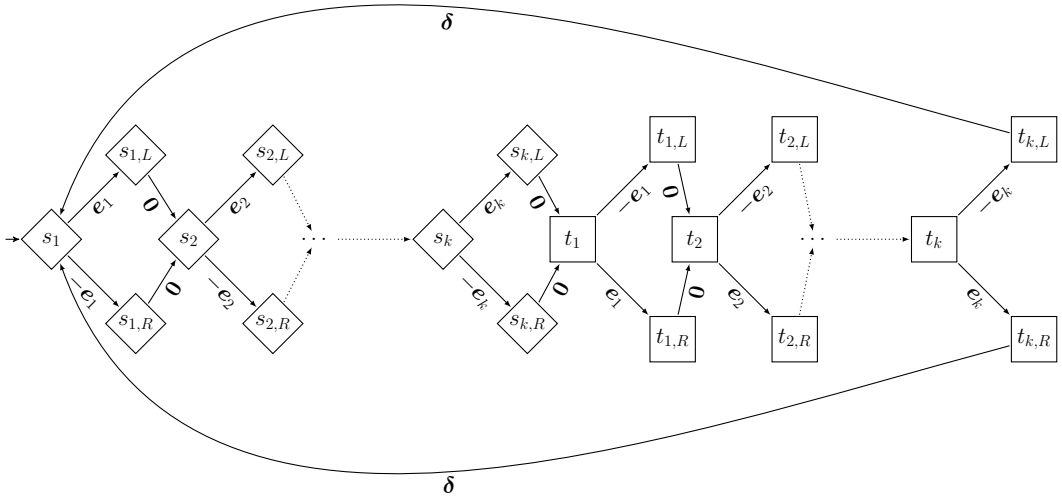

Figure 5.1: In the game $\mathcal{G}_k$, optimal Max strategies for $\mathtt{MP} > \mathbf{0}$ for $2k$ dimensions require at least 2^k memory modes. We have $2k$-dimensional reward vectors $\boldsymbol{e_i}$ such that $\boldsymbol{e_i}[2i-1] = +1$, $\boldsymbol{e_i}[2i] = -1$ and 0 elsewhere. The special vector $\boldsymbol{\delta} = (\delta, \ldots, \delta)$ has value $\delta \stackrel{\text{def}}{=} 2^{-2k}$ in every dimension.

▶ **Theorem 5.2.** *There is a family of deterministic games $\mathcal{G}_k$ with $6k$ states as in Figure 5.1, such that every state is almost surely winning for the $2k$-dimensional $\mathtt{MP} > \mathbf{0}$ objective for Max, but any randomized Max strategy with $< 2^k$ memory modes cannot win almost surely.*

Proof. Consider the games $\mathcal{G}_k$ from Figure 5.1 with $6k$ states. The dimension of the rewards $2k$ is split into k blocks of 2 dimensions each and the reward vector $\boldsymbol{e_i}$ is $(+1, -1)$ on the i-th block and zero elsewhere. First Min makes k decisions, choosing between $\boldsymbol{e_i}$ and $-\boldsymbol{e_i}$ for each $i = 1, 2, \ldots, k$. Then Max makes k decisions, choosing between $-\boldsymbol{e_i}$ and $\boldsymbol{e_i}$ for each $i = 1, 2, \ldots, k$. Max can win $\mathtt{MP} > \mathbf{0}$ surely (from s_1 and thus from every other state) by exactly copying Min's choices such that these rewards cancel out, which leaves just the strictly positive reward vector $\boldsymbol{\delta} = (2^{-2k}, \ldots, 2^{-2k}) > \mathbf{0}$ between each visit to s_1. (Since $\boldsymbol{\delta}$ can be described with a number of bits that is polynomial in k, $\mathcal{G}_k$ has polynomial size.)

Intuitively, Max needs at least 2^k memory modes to remember which of the 2^k possible choices Min made, in order to copy it. While this is easy to prove for deterministic Max strategies, we show a stronger result: Even randomized Max

strategies with $< 2^k$ memory modes cannot win almost surely.

Let σ be an arbitrary randomized Max strategy with $m < 2^k$ memory modes. Overall, for the k choices in the states $s_1, \ldots, s_k$, Min has 2^k different possible options, which we denote by index numbers $j \in \{1, \ldots, 2^k\}$. (Similarly for Max's options in states $t_1, \ldots, t_k$.) Let π_j be the deterministic Min strategy that always plays option j forever. Let π be the randomized Min strategy which initially picks a $j \in \{1, \ldots, 2^k\}$, with equal probability 2^{-k}, and then plays π_j forever. Now we show that $\mathcal{P}^{\mathcal{G}_k}_{\sigma,\pi,s_1}(\mathtt{MP} > \mathbf{0}) < 1$.

Let M^j_k be the Markov chain derived from $\mathcal{G}_k$ by fixing σ and π_j. If M^j_k is irreducible and aperiodic then let P be the $2^k \times m$ matrix such that $P(j, i)$ is the probability, in the steady-state, that σ is in memory mode $i \in \{1, \ldots, m\}$ in state t_1. (It is also possible that M^j_k is reducible, since it has a memory component from the strategy σ as part of its state. If σ is such that it would never visit certain memory modes again eventually, then this makes M^j_k reducible. In this case, we consider an irreducible subchain. If M^j_k is periodic then consider the long-run average frequency of being in memory mode i in state t_1 instead of the probability in the following arguments.) Let Q be the $m \times 2^k$ matrix where $Q(i, j)$ is the probability that σ will play option j in states $t_1, \ldots, t_k$ if it is in memory mode i in state t_1.

The probability that Max will play option j in M^j_k in the long run is then given by $(PQ)(j, j)$, the j-th diagonal value in the product matrix. Since $m < 2^k$, the $2^k \times 2^k$ matrix PQ does not have full rank. Thus at least one of its eigenvalues is 0. Moreover, since both matrices P and Q are row-stochastic, the absolute value of the eigenvalues of PQ are upper-bounded by 1. So the sum of the eigenvalues of PQ is upper-bounded by $2^k - 1$. The Cayley–Hamilton theorem implies that the trace is the sum of the eigenvalues. This yields an upper bound on the trace of PQ, namely $Tr(PQ) = \sum_{j=1}^{2^k} (PQ)(j, j) \leq 2^k - 1$. Hence there exists a j' such that $(PQ)(j', j') \leq 1 - 2^{-k}$. So, in $M^{j'}_k$, the long-run probability that Max will pick some option different from j' is at least 2^{-k}. Hence there exists at least one option $j'' \neq j'$ that is picked by Max in the long run with probability $\geq 2^{-k}2^{-k} = 2^{-2k}$. Since $j'' \neq j'$, there exists at least one block of two dimensions, say x and $x + 1$, where Max choosing j' has effect $(+1, -1)$ and j'' has the opposite effect $(-1, +1)$ (or vice-versa; this case is symmetric). Thus, between two visits to state s_1, the expected reward in dimension x is $\leq -2 \cdot 2^{-2k} + \delta = -2^{-2k} < 0$. Therefore, in $M^{j'}_k$, the $\mathtt{MP} > \mathbf{0}$ objective is satisfied with probability zero, since it almost surely

fails in dimension x. I.e., $\mathcal{P}^{M_k^{j'}}_{\sigma,\pi_{j'},s_1}(\mathtt{MP} > \mathbf{0}) = 0$.

Since Min's strategy π initially decides to become $\pi_{j'}$ with probability 2^{-k}, we obtain that $\mathcal{P}^{\mathcal{G}_k}_{\sigma,\pi,s_1}(\mathtt{MP} > \mathbf{0}) \leq 1 - 2^{-k}(1 - \mathcal{P}^{M_k^{j'}}_{\sigma,\pi_{j'},s_1}(\mathtt{MP} > \mathbf{0})) = 1 - 2^{-k} < 1$. ◀

5.4 Strategy Complexity of $\mathtt{MP} > 0 \cap \mathtt{EPAR}$ With Randomization

In this section, unless otherwise stated, we consider the one-dimensional $\mathtt{MP} > 0 \cap \mathtt{EPAR}$ objective, where Max needs to attain a strictly positive one-dimensional mean payoff while also satisfying a max-even parity condition. These are defined via a reward function r and a coloring function Col. W.l.o.g. we assume that the range of r is $\subseteq [-R, R]$ for some $R > 0$ and the colors are either $\{0, 1, \ldots, d\}$ or $\{1, \ldots, d\}$.

For almost surely winning strategies in MDPs, *deterministic* Max strategies require $\Theta(\exp(\|\mathcal{M}\|))$ memory modes, *i.e.*, exponential memory is both necessary and sufficient [GOP11, Theorem 5].

We show that *randomized* Max strategies require less memory: none in MDPs and just polynomially many memory modes in stochastic games.

▶ **Theorem 5.3.** *In maximizing MDPs, almost surely winning strategies for the multi-dimensional* $\mathtt{MP} > \mathbf{0} \cap \mathtt{EPAR}$ *objective can be chosen memoryless randomized (MR).*

Proof. We use the fact that, for any strategy, almost surely all the runs eventually end up in an end component [DA97, Theorem 3.2]. In a finite-state MDP, there can only be a finite (albeit exponential) number of end components. Let $\mathcal{M}_1 \stackrel{\text{def}}{=} \left(S_1, S^1_\Box, S^1_\bigcirc, E_1, P_1\right)$ be one such end component of $\mathcal{M}$. We say that $\mathcal{M}_1$ is 'winning' iff the highest color is even and $S_1 \subseteq \mathtt{AS}^{\mathcal{M}_1}_\Box(\mathtt{MP} > \mathbf{0})$. Given an almost surely winning strategy σ^*, except for a null set, all induced runs must eventually stay inside a winning end component. Thus from every almost surely winning state it is possible to almost surely reach a winning end component.

▷ Claim 5.4. Let $\mathcal{M}_1$ be a winning end component. Then there is a memoryless randomized strategy σ^* which is almost surely winning for $\mathtt{MP} > \mathbf{0} \cap \mathtt{EPAR}$ from every state in $\mathcal{M}_1$.

Proof. By definition of a winning end component, we know that $S_1 \subseteq \mathtt{AS}_{\Box}^{\mathcal{M}_1}(\mathtt{MP} > \mathbf{0})$. From the proof of [BBC+14, Prop. 5.1 (ii)] we obtain that there exists a memoryless randomized strategy ξ_ε for some $\varepsilon > 0$, such that ξ_ε almost surely wins $\mathtt{MP} > \mathbf{0}$ from any state in S_1 and $\|\varepsilon\|$, the size of the probabilities used by ξ_ε, is polynomial in $\|\mathcal{M}_1\|$. Furthermore, every edge in E_1 is used with some positive probability. This implies that ξ_ε also satisfies $\mathtt{EPAR}$ almost surely, since the highest color in $\mathcal{M}_1$ is even. ◁

Consider the union of all winning end components $\mathcal{C}$ in $\mathcal{M}$. We obtain an almost surely winning MR strategy for $\mathtt{MP} > \mathbf{0} \cap \mathtt{EPAR}$ from every state in $\mathcal{M}$ as follows. Outside of $\mathcal{C}$ it plays an almost surely winning uniform MD strategy for the reachability objective $\mathsf{F}\,\mathcal{C}$. Inside each maximal winning end component $\mathcal{M}_1$ in $\mathcal{C}$ it plays the MR strategy ξ_ε from Claim 5.4.

◀

Although almost surely winning $\mathtt{MP} > 0 \cap \mathtt{EPAR}$ has not been studied in the case of stochastic games, since both $\mathtt{MP} > 0$ and $\mathtt{EPAR}$ are shift-invariant and inverse-submixing ([BBE10b, Lemma 6] for $\mathtt{MP} > 0$, [GK23, Prop. 3.1]), the same holds for the conjunction $\mathtt{MP} > 0 \cap \mathtt{EPAR}$. Applying Theorem 5.1 ([GK23, Sec. 6.1]), implies that the memory required by Max in stochastic games is also $\Theta(\exp(\|\mathcal{G}\|))$ when considering deterministic strategies.

For general strategies, even with the improved upper bound in MDPs of Theorem 5.3, the construction in Theorem 5.1 ([GK23, Sec. 6.1]) still only yields an exponential upper bound on the memory of Max strategies in stochastic games. We show that this is excessive in general. Optimal Max strategies for $\mathtt{MP} > 0 \cap \mathtt{EPAR}$ in stochastic games require, at least and at most, #(distinct even colors) many memory modes (*i.e.*, polynomial memory).

▶ **Theorem 5.5.** *Consider a game $\mathcal{G} = (S, (S_\Box, S_\Diamond, S_\bigcirc), E, P)$, coloring function Col with highest color d, reward function r and objective $\mathtt{MP} > 0 \cap \mathtt{EPAR}$ for player Max.*

1. *If Max can win almost surely from some state s then there also exists an almost surely winning randomized Max strategy from s that uses at most k memory modes and total bit size $\mathcal{O}\left(\|\mathcal{G}\|^{d+c}\right)$ where $k = \#(\text{distinct even colors})$, d the highest color, and c is a constant independent of $\mathcal{G}$.*

2. *The bound on the number of required memory modes for almost surely winning randomized Max strategies is tight. There is a family of deterministic games* $\{\mathcal{G}_n \mid n \geq 2\}$ *as in Definition 5.18 and Figure 5.2 such that*

   - $\|\mathcal{G}_n\| = \Theta(n)$, $\mathcal{G}_n$ *contains* n *even colors and every state is almost surely winning for Max.*
   - *For every* $n \geq 2$*, any Max strategy with* $< n$ *memory modes is worthless,* i.e.*, it cannot guarantee* $\texttt{MP} > 0 \cap \texttt{EPAR}$ *with any positive probability.*

The idea of using randomization to reduce memory complexity is not new and appears, e.g., in [CDGH15, Cha07, Hor09]. However, it is interesting to note that the Max strategy still requires a small amount of memory even in the presence of randomization and cannot be made memoryless, unlike in MDPs (Theorem 5.3)

Strategy \ Arena	MDPs $\mathcal{M}$	Stochastic games $\mathcal{G}$
Deterministic	$\Theta(\exp(\|\mathcal{M}\|))$ [GOP11]	$\Theta(\exp(\|\mathcal{G}\|))$ [GOP11, GK23]
Randomized	1	$\#(\text{distinct even colors}) = \Theta(\|\mathcal{G}\|)$

Table 5.1: Worst case number of memory modes required for almost surely winning Max strategies for objective $\texttt{MP} > 0 \cap \texttt{EPAR}$ in MDPs and stochastic games.

Table 5.1 highlights the strategy complexity of almost surely winning Max strategies in MDPs and games for the $\texttt{MP} > 0 \cap \texttt{EPAR}$ objective. Note that while there is no explicit result for the complexity of deterministic winning strategies, the results from [GOP11] and lifting this strategy using Theorem 5.1 ([GK23, Sec. 6.1]) provide tight bounds up-to a constant. The entries in the randomized row are our contributions. The bit size of these randomized Max strategies is $\mathcal{O}(\|\mathcal{M}\|^c)$ and $\mathcal{O}\left(\|\mathcal{G}\|^{2k+c_1}\right)$ in MDPs and games respectively, where $k = \#(\text{distinct even colors})$ and c and c_1 are constants.

We sketch the main idea of the proof of Item 1.. It follows the induction argument of [CDGO14, Lemma 2,3] along with strategy complexity analysis for $\texttt{MP} > 0 \cap \texttt{EPAR}$ instead of $\texttt{MP} \geq 0 \cap \texttt{EPAR}$. There are two cases, depending on whether the maximum color in the game is odd or even.

If the maximum color d is odd, then the EPAR part of $\texttt{MP} > 0 \cap \texttt{EPAR}$ requires that this color must eventually never been seen any more (except in a null set of

the plays). Using a ranking argument, we show that it is possible to partition the state space into layers $Z_1, \cdots, Z_\ell$ such that

- Max can force the win in any one of these layers if the play stays there forever.
- Min cannot infinitely often switch between the layers, except in a null set of the plays.

The above two facts can be combined to build a winning Max strategy. Interestingly, this strategy does not use any additional memory, compared to the inductive case with one less color. It can be seen as a combination of memoryless attractor strategies on certain states, combined with almost surely winning strategies in other states in subgames with at least one less color (using the IH).

If the maximum color d is even, Max tries wherever possible to reach a state of this color but does so sufficiently infrequently, so that this does not compromise the satisfaction of the positive mean-payoff objective. This is achieved by operating in phases with Max playing either for $\texttt{MP} > 0 \cap \mathsf{F}(S(d))$ or for $\texttt{MP} > 0 \cap \texttt{EPAR}$ in a subgame with at least 1 less color. Let S denote the set of states in the game $\mathcal{G}$, and $X \stackrel{\text{def}}{=} \texttt{Attr}_\Box(S(d))$ the positive attractor of states with color d in $\mathcal{G}$, $Y \stackrel{\text{def}}{=} S \setminus X$. Since $\mathcal{G}[Y]$ has at least 1 less color, by induction hypothesis let σ^*_Y denote the almost surely winning strategy for Max in $\mathcal{G}[Y]$. Let $\sigma^*_{\texttt{MP}}$ denote the optimal MD strategy for $\texttt{MP} > 0$ in $\mathcal{G}$ and $\sigma^*_{\texttt{Attr}}$ denote the positive attractor strategy in states in $X \cap S_\Box$. Then the optimal randomized Max strategy σ^* works as follows.

Phase-1 If the current state is in X, play the mixed strategy $\varepsilon_0\sigma^*_{\texttt{Attr}} + (1-\varepsilon_0)\sigma^*_{\texttt{MP}}$. I.e., play $\sigma^*_{\texttt{Attr}}$ with some sufficiently small probability $\varepsilon_0 > 0$ and play $\sigma^*_{\texttt{MP}}$ with probability $1-\varepsilon_0$. Else, play according to $\sigma^*_{\texttt{MP}}$. At every step, there is a sufficiently small chance $\varepsilon_1 > 0$ to stop this phase. This is done by changing Max's memory mode with probability ε_1. So Phase-1 stops eventually almost surely, where the expected time to stopping depends on ε_1. This serves as a makeshift probabilistic clock (since this strategy does not use a real clock). After this phase is over, if the play is in X, restart Phase-1, else move to Phase-2.

Phase-2 Play according to σ^*_Y while the play is in Y. If the play ever moves to X, switch to Phase-1.

One has to choose ε_0, $\varepsilon_1 > 0$ so that the overall strategy is almost surely winning. Intuitively, $\varepsilon_0 > 0$ is chosen so small that the mean-payoff is strictly positive in Phase-1, but this holds only in the long run. Moreover, while Phase-2 also yields a strictly positive mean-payoff in the long run, it might switch back to Phase-1 early while the accumulated reward is still negative. (However, this possible negative reward can be bounded, in expectation.) Therefore $\varepsilon_1 > 0$ is chosen so small that Phase-1 is played for a very long time (in expectation). Hence Phase-1 closely approximates its long run behavior with strictly positive mean-payoff, and thus it also compensates for any possible negative reward obtained during a temporary switch to Phase-2.

Also note that this strategy uses randomization in both the memory updates and the next move. Thus it is an FRR strategy. Furthermore, the strategy needs 1 additional memory mode (compared to strategy σ^*_Y) in order to know the current phase. This is necessary, because it needs to play different strategies in Y: $\sigma^*_{\texttt{MP}}$ in Phase-1 and σ^*_Y in Phase-2.

Proof of Item 1.(Theorem 5.5). The proof is by strong induction on the maximum color d.
Let P_d denote the statement as given by Item 1.. We then prove the following.

$$P_0(\text{if minimum color is 0}) \tag{5.1}$$

$$P_1(\text{if minimum color is 1}) \tag{5.2}$$

$$k \geq 0\ (\forall\ s \leq 2k.\ P_s) \implies P_{2k+1} \tag{5.3}$$

$$k \geq 0\ (\forall\ s \leq 2k+1.\ P_s) \implies P_{2k+2} \tag{5.4}$$

Base case: When there is only one color, `EPAR` is trivial(Equation (5.1)) or never satisfied(Equation (5.2)) and hence $\texttt{MP} > 0 \cap \texttt{EPAR}$ is equivalent to $\texttt{MP} > 0$ for which MD strategies exist [BBE10a, Prop. 7]. Since d is either 0 or 1, either the number of even colors is 1 or no almost surely winning strategy exists for Max.
Induction Step: Assume that the statement is true for all $k < d$. W.l.o.g. one can assume that every state s is almost surely winning as otherwise it is possible to consider a subgame with states that satisfy this condition. To show that statement holds for d, we split into two cases depending on whether d is even or odd.

▷ Claim 5.6 (Maximum Color Odd). Let $d = 2k+1$ be odd. Then there is an almost surely winning Max strategy σ^* which can be constructed from a finite number of σ^*_i which are finite-memory almost surely winning strategies in subgames with

maximum color $< d$. Moreover, the number of memory modes in σ^* is equal to the maximum number of memory modes in σ_i^*. If all σ_i^* have rational probabilities, then $\text{bits}(\sigma^*) = \mathcal{O}(\sum_i \text{bits}(\sigma_i^*)) = \mathcal{O}(|S| \text{ bits}(\sigma_i^*))$

Before going to the proof, we first prove a simple claim which we need.

▷ Claim 5.7 (Positive Attractor property). In a game $\mathcal{G} = (S, (S_\Box, S_\Diamond, S_\bigcirc), E, P)$, a subset of the states $H \subseteq S$, and a finite-memory Max strategy σ^* that satisfies the following condition: $\forall s \in \mathtt{Attr}_\Box(H)\, \forall \mathsf{m}.\, \inf_\pi \mathcal{P}_{\sigma^*[\mathsf{m}],\pi,s}(\mathsf{F}\, H) > 0$. Then the following two properties hold.

1. $\mathcal{P}_{\sigma^*,\pi,s_0}(\mathsf{GF}\, H \,\Delta\, \mathsf{GF}\,(\mathtt{Attr}_\Box\, H)) = 0$.
2. Furthermore, if H is a Min-trap and σ^* always remains inside $\mathcal{G}[H]$ after entering H, then $\mathcal{P}_{\sigma^*,\pi,s_0}(\mathsf{FG}\, H \,\Delta\, \mathsf{GF}\, H) = 0$.

Proof of Claim 5.7. Towards Item 1., we use the facts that $\mathtt{Attr}_\Box(H) \subseteq S$ is finite and σ^* is finite-memory to obtain that $p \stackrel{\text{def}}{=} \min_{s \in \mathtt{Attr}_\Box(H)} \min_{\mathsf{m}} \inf_\pi \mathcal{P}_{\sigma^*[\mathsf{m}],\pi,s}(\mathsf{F}\, H) > 0$. Hence $\mathcal{P}_{\sigma^*,\pi,s_0}(\mathsf{GF}\, H \,\Delta\, \mathsf{GF}\,(\mathtt{Attr}_\Box\, H)) = \mathcal{P}_{\sigma^*,\pi,s_0}(\mathsf{GF}\,(\mathtt{Attr}_\Box\, H) \setminus \mathsf{GF}\, H) \leq (1-p)^\infty = 0$.

Item 2. follows directly from the assumptions. ◁

It is easy to see that the above claim also holds true from the perspective of Min, *i.e.*, when Max and Min roles are reversed.

We now turn to the proof of Claim 5.6.

Proof. of Claim 5.6. For Max to win $\mathtt{MP} > 0 \cap \mathtt{EPAR}$ almost surely, it has to eventually not see any of the states from $S(d)$ any more. To achieve this, we 'rank' the states in $S(d)$ so that it is not possible for Min to force a move from a lower ranked state to a higher ranked one. Ultimately, this implies that eventually almost surely we are always inside states with the same rank. If we show that Max can win here, then we are done. Formalizing this intuition, we need to show that

1. There exists a partition $\{Z_i\}_{1 \leq i \leq \ell}$ of S and non-empty sets R_i, U_i for $1 \leq i \leq \ell$ where $U_1 = S$, and $U_{\ell+1} = \emptyset$ such that
   (a). $R_i \subseteq U_i \setminus U_i(d) \neq \emptyset$ is a trap for Min in $\mathcal{G}[U_i]$ and $R_i \subseteq \mathtt{AS}_\Box^{\mathcal{G}[R_i]}(\mathtt{MP} > 0 \cap \mathtt{EPAR})$ with corresponding winning strategy σ_i^*
   (b). $Z_i = \mathtt{Attr}_\Box(R_i, \mathcal{G}[U_i])$

(c). $U_{i+1} = U_i \setminus Z_i$

The states in Z_i as defined above for $1 \leq i \leq \ell$ can be thought of as having rank i. Note that finding a non-empty R_i in U_i fixes Z_i and U_{i+1}. The above characterization can be seen as inductively defining the sets Z_i as long as one can find the non-empty set R_i. The induction stops when U_{i+1} is empty. The R_i is not unique and different R_i's thus give rise to different partitions. The claim holds for any partition satisfying the above conditions. To explicitly find one such non-empty R_i, we can use the following claim.

▷ Claim 5.8. Let the maximum color d be odd and all states $\in \mathtt{AS}_{\Box}(\mathtt{MP} > 0 \cap \mathtt{EPAR})$. Let $\emptyset \subset U \subseteq S$ be a trap for player Max. Define $X_U \stackrel{\text{def}}{=} \mathtt{Attr}_{\Diamond}(U(d), \mathcal{G}[U])$ and $Y_U \stackrel{\text{def}}{=} U \setminus X_U$. Then

$$R_U \stackrel{\text{def}}{=} \mathtt{AS}_{\Box}^{\mathcal{G}[Y_U]}(\mathtt{MP} > 0 \cap \mathtt{EPAR}) \neq \emptyset.$$

Proof of Claim 5.8. Towards a contradiction, assume $R_U = \emptyset$. We'll show that this implies that there is a Min strategy $\pi = (\mathsf{M}, \mathsf{m}_0, \mathsf{upd}, \mathsf{nxt})$ always staying in Y_U such that for all states $s \in Y_U$ and all strategies σ of player Max which always stay in Y_U,

$$\mathcal{P}^{\mathcal{G}[Y_U]}_{\sigma,\pi[\mathsf{m}_0],s}(\mathtt{MP} > 0 \cap \mathtt{EPAR}) < 1.$$

Let $E \stackrel{\text{def}}{=} \mathsf{F}\,\mathsf{G}\,Y_U$ and $F \stackrel{\text{def}}{=} \overline{E}$ in $\mathcal{G}[U]$. Then, $F = \mathsf{G}\,\mathsf{F}\,X_U$. By definition of X_U, there is a MD Min strategy $\pi_{\mathtt{Attr},U}$ from every Min state in X_U. Consider the following Min strategy $\pi^* \stackrel{\text{def}}{=} (\mathsf{M}', \mathsf{upd}', \mathsf{nxt}')$ where

- $\mathsf{M}' \stackrel{\text{def}}{=} \mathsf{M}$
- $\mathsf{nxt}'(\mathsf{m}, s) \stackrel{\text{def}}{=} \begin{cases} \pi_{\mathtt{Attr},U}(s) & s \in X_U \\ \mathsf{nxt}(\mathsf{m}, s) & s \in Y_U \end{cases}$
- $\mathsf{upd}'(\mathsf{m}, (s, s')) \stackrel{\text{def}}{=} \begin{cases} \mathsf{m}_0 & s \text{ or } s' \in X_U \\ \mathsf{upd}(\mathsf{m}, (s, s')) & \text{otherwise} \end{cases}$

Observe that $\pi^*[\mathsf{m}_0]$ never leaves U. Consider any Max strategy σ, start state

$s \in U$. Then $\mathcal{P}^{\mathcal{G}}_{\sigma,\pi^*[\mathsf{m}_0],s}(\mathtt{MP} > 0 \cap \mathtt{EPAR})$

$$
\begin{aligned}
&= \mathcal{P}^{\mathcal{G}}_{\sigma,\pi^*[\mathsf{m}_0],s}(\mathtt{MP} > 0 \cap \mathtt{EPAR} \cap E) && + \mathcal{P}^{\mathcal{G}}_{\sigma,\pi^*[\mathsf{m}_0],s}(\mathtt{MP} > 0 \cap \mathtt{EPAR} \cap F)\\
&= \mathcal{P}^{\mathcal{G}}_{\sigma,\pi^*[\mathsf{m}_0],s}(E) \cdot \underbrace{\mathcal{P}^{\mathcal{G}[Y_U]}_{\sigma,\pi^*[\mathsf{m}_0],\alpha}(\mathtt{MP} > 0 \cap \mathtt{EPAR})}_{<1} && + \mathcal{P}^{\mathcal{G}}_{\sigma,\pi^*[\mathsf{m}_0],s}(\mathtt{MP} > 0 \cap \mathtt{EPAR} \cap F)\\
&< \mathcal{P}^{\mathcal{G}}_{\sigma,\pi^*[\mathsf{m}_0],s}(E) && + \underbrace{\mathcal{P}^{\mathcal{G}}_{\sigma,\pi^*[\mathsf{m}_0],s}(\mathtt{EPAR} \cap \mathsf{GF}\, U(d))}_{\text{by } \textit{Claim } 5.7}\\
&< \mathcal{P}^{\mathcal{G}}_{\sigma,\pi^*[\mathsf{m}_0],s}(E) && + \underbrace{0}_{d \text{ is the max color and odd}}\\
&< \mathcal{P}^{\mathcal{G}}_{\sigma,\pi^*[\mathsf{m}_0],s}(E)\\
&< 1
\end{aligned}
$$

which implies $U \not\subseteq \mathtt{AS}_{\Box}(\mathtt{MP} > 0 \cap \mathtt{EPAR})$ a contradiction. ◁

Since $U_1 = S$ itself is a trap for Max, Claim 5.8 implies existence of a non-empty set R_1 such that $R_1 \subseteq U_1 \setminus U_1(d)$ and $R_1 \subseteq \mathtt{AS}^{\mathcal{G}[R_1]}_{\Box}(\mathtt{MP} > 0 \cap \mathtt{EPAR})$ (true since a winning strategy never leaves the almost sure set of states). Define $Z_1 \stackrel{\text{def}}{=} \mathtt{Attr}_{\Box}(R_1, \mathcal{G}[U_1])$ and let $U_2 \stackrel{\text{def}}{=} U_1 \setminus Z_1$. Observe that U_2 is a strict subset of U_1 and is once again a trap for Max. If the highest color in U_2 is d, one can again apply Claim 5.8 or else trivially take $R_2 \stackrel{\text{def}}{=} U_2$ and the procedure stops.

For our purposes, any partition which satisfies Item 1. suffices for the proof. Observe that $U_i = \bigcup_{k=i}^{\ell} Z_k$ and $U_i \supset U_{i+1}$. Also $R_i \subseteq Z_i$ and $R_i \subseteq U_i \setminus U_i(d)$ implies $R_i \subseteq Z_i \setminus Z_i(d)$. Furthermore, given $\sigma^*_i[\mathsf{m}_0] = (\mathsf{M}, \mathsf{nxt}_i, \mathsf{upd}_i)$ which is winning from every state in R_i and the uniform MD attractor strategies $\sigma^*_{\mathtt{Attr},i}$ to R_i in Z_i, we construct a strategy $\sigma^*[\mathsf{m}_0] \stackrel{\text{def}}{=} (\mathsf{M}, \mathsf{nxt}, \mathsf{upd})$ as follows.

$$
\mathsf{nxt}(\mathsf{m}, s) \stackrel{\text{def}}{=} \begin{cases} \sigma^*_{\mathtt{Attr},i}(s) & s \in Z_i \setminus R_i \text{ for some } 1 \le i \le \ell \\ \mathsf{nxt}_i(\mathsf{m}, s) & s \in R_i \text{ for some } 1 \le i \le \ell \end{cases}
$$

$$
\mathsf{upd}(\mathsf{m}, (s, s')) \stackrel{\text{def}}{=} \begin{cases} \mathsf{m}_0 & s \in Z_i \setminus R_i,\ s' \in Z_i \text{ for some } 1 \le i \le \ell \\ \mathsf{upd}_i(\mathsf{m}, (s, s')) & s,\ s' \in R_i \text{ for some } 1 \le i \le \ell \\ \mathsf{m}_0 & s \in Z_i,\ s' \in Z_j\ i \ne j \end{cases}
$$

We assumed that all the strategies σ^*_i share the same set of memory configurations M and start in the same memory configuration m_0. If they are different, one can take M such that $|\mathsf{M}| = \max_i(|\mathsf{M}_i|)$ and simple renaming of the configurations in the strategies so that every start configuration is renamed to m_0. From the definition,

it is clear that $\|\sigma^*\| = \max_i \|\sigma_i^*\|$ and $\text{bits}(\sigma^*) = \sum_i \text{bits}(\mathsf{nxt}_i) + \text{bits}(\mathsf{upd}_i) = \mathcal{O}(|S| \text{bits}(\sigma_i^*))$. This takes care of the complexity part of the claim. We now show that $\sigma^*[\mathsf{m}_0]$ is almost surely winning from every state s.

Let $\mathtt{O}_i$ denote $\mathsf{F}\,\mathsf{G}\,R_i$ for each $1 \leq i \leq \ell$ and consider the event $\mathtt{MP} > 0 \cap \mathtt{EPAR} \cap \mathtt{O}_i$. Fix some arbitrary strategy π for Min and assume $\mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}(\mathtt{O}_i) > 0$ as otherwise the conclusion obtained below is trivial. For every play $\rho \in \mathtt{O}_i$, let the random variable T_i denote the hitting time of a state which satisfies $\mathsf{G}\,R_i$ and X_{T_i} denote the state in which one enters R_i. Also, let α_i denote the distribution of X_{T_i} and $\pi_i \in \Pi_{\mathcal{G}[R_i]}$ denote the Min strategy which simulates the play until $\mathsf{G}\,R_i$ in $\mathcal{G}$ and then plays according to π. Then

$$\mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}(\mathtt{MP} > 0 \cap \mathtt{EPAR} \cap \mathtt{O}_i) = \mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}(\mathtt{MP} > 0 \cap \mathtt{EPAR} \mid \mathtt{O}_i) \cdot \mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}(\mathtt{O}_i)$$

$$\begin{aligned}
&= \sum_{s' \in R_i} \mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}(\mathtt{MP} > 0 \cap \mathtt{EPAR} \mid X_{T_i} = s') \cdot \mathcal{P}^{\mathcal{G}}(X_{T_i} = s' \mid \mathtt{O}_i) \cdot \mathcal{P}^{\mathcal{G}}(\mathtt{O}_i) \\
&= \mathcal{P}^{\mathcal{G}[R_i]}_{\sigma_i^*[\mathsf{m}_0],\pi_i,\alpha_i}(\mathtt{MP} > 0 \cap \mathtt{EPAR}) \cdot \mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}(\mathtt{O}_i) \\
&= \mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}(\mathtt{O}_i)
\end{aligned}$$

From the above discussion, one can see that $\mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}(\mathtt{MP} > 0 \cap \mathtt{EPAR} \cap \bigcup_i \mathtt{O}_i) = \mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}(\bigcup_i \mathtt{O}_i)$. If we show the latter event occurs with probability 1, then we are done with the converse since

$$\mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}(\mathtt{MP} > 0 \cap \mathtt{EPAR}) \geq \mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}\left(\mathtt{MP} > 0 \cap \mathtt{EPAR} \cap \bigcup_i \mathtt{O}_i\right) = 1.$$

Consider the complement objective and its probability $\mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}\left(\overline{\bigcup_i \mathtt{O}_i}\right)$ which can equivalently be written as $\mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}\left(\bigcap_i \overline{\mathtt{O}_i}\right)$. Z_i is the attractor for R_i within $\mathcal{G}[U_i]$ where R_i is also a Min-Trap. $\sigma^*[\mathsf{m}_0]$ when in $\mathcal{G}[U_i]$ satisfies the hypothesis given in Claim 5.7. Therefore for every i,

$$\mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}((\mathsf{G}\,\mathsf{F}\,Z_i \,\Delta\, \mathsf{G}\,\mathsf{F}\,R_i) \mid \mathsf{F}\,\mathsf{G}\,U_i) = 0 \tag{5.5}$$

$$\mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}((\mathsf{G}\,\mathsf{F}\,R_i \,\Delta\, \mathtt{O}_i) \mid \mathsf{F}\,\mathsf{G}\,U_i) = 0 \tag{5.6}$$

From (5.5), (5.6), one can then conclude

$$\begin{aligned}
\mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}\big(\mathsf{F}\,\mathsf{G}\,U_i \cap \overline{\mathtt{O}_i}\big) &= \mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}\big(\mathsf{F}\,\mathsf{G}\,U_i \cap \overline{\mathsf{G}\,\mathsf{F}\,R_i}\big) \\
&= \mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}\big(\mathsf{F}\,\mathsf{G}\,U_i \cap \overline{\mathsf{G}\,\mathsf{F}\,Z_i}\big) \\
&= \mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}\big(\mathsf{F}\,\mathsf{G}\,U_i \cap \mathsf{F}\,\mathsf{G}\,\overline{Z_i}\big) \\
&= \mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}\big(\mathsf{F}\,\mathsf{G}\,\big(U_i \cap \overline{Z_i}\big)\big) \\
&= \mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}(\mathsf{F}\,\mathsf{G}\,U_{i+1}).
\end{aligned}$$

Since $U_1 = S$, $\mathsf{F}\,\mathsf{G}\,U_1$ is a sure event and hence

$$\begin{aligned}
\mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}\left(\bigcap_i \overline{\mathtt{O}_i}\right) &= \mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}\left(\mathsf{F}\,\mathsf{G}\,U_1 \cap \overline{\mathtt{O}_1} \cap \bigcap_{i>1} \overline{\mathtt{O}_i}\right) \\
&= \mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}\left(\mathsf{F}\,\mathsf{G}\,U_2 \cap \overline{\mathtt{O}_2} \cap \bigcap_{i>2} \overline{\mathtt{O}_i}\right) \\
&= \mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}\left(\mathsf{F}\,\mathsf{G}\,U_j \cap \overline{\mathtt{O}_j} \cap \bigcap_{i>j} \overline{\mathtt{O}_i}\right) \\
&= \mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}(\mathsf{F}\,\mathsf{G}\,U_{l+1}) \\
&= 0
\end{aligned}$$

◁

Remark 5.9. It is easy to see that the only property of $\mathtt{MP} > 0$ used in the proof was that it is shift-invariant. Hence, the above claim can be generalized to any objective of the form $\mathtt{O} \cap \mathtt{EPAR}$ with $\mathtt{O}$ being shift-invariant. However, the claim is presented in its present form since this observation by itself is of little significance if one cannot generalize the case with the highest color being even as well to utilize the induction argument.

▷ Claim 5.10 (Maximum Color Even). Let $d > 0$ be even and assume there are k even colors. Define $X \stackrel{\text{def}}{=} \mathtt{Attr}_{\Box}(S(d), \mathcal{G})$ and $Y \stackrel{\text{def}}{=} S \setminus X$.

1. $S \subseteq \mathtt{AS}^{\mathcal{G}}_{\Box}(\mathtt{MP} > 0 \cap \mathtt{EPAR})$ if and only if
   - $S \subseteq \mathtt{AS}^{\mathcal{G}}_{\Box}(\mathtt{MP} > 0)$ and
   - $Y \subseteq \mathtt{AS}^{\mathcal{G}[Y]}_{\Box}(\mathtt{MP} > 0 \cap \mathtt{EPAR})$

2. Let σ^*_Y denote a Max strategy in $\mathcal{G}[Y]$, which has finite memory and is almost surely winning from every state in Y. Then an almost surely winning Max strategy σ^* in $\mathcal{G}$ can be constructed such that $\|\sigma^*\| = 1 + \|\sigma^*_Y\|$. Furthermore, if all the probabilities in σ^*_Y are rational, then $\text{bits}(\sigma^*) = \mathcal{O}(\|\mathcal{G}\|^{c_1}\|\sigma^*_Y\| + \text{bits}(\sigma^*_Y)\|\mathcal{G}\|)$ where c_1 is a fixed constant independent of $\mathcal{G}$.

Proof of Claim 5.10. For Item 1., one direction (forward) is trivial. For the other direction (converse), by our assumptions, there exist

1. A uniform almost surely winning Max strategy $\sigma_{\mathtt{MP}}$ that is MD such that against any Min strategy π and from any start state s,

$$\mathcal{P}^{\mathcal{G}}_{\sigma_{\mathtt{MP}},\pi,s}(\mathtt{MP} > 0) = 1$$

2. A almost surely winning Max strategy $\sigma_Y \stackrel{\text{def}}{=} (\mathsf{M}, \mathsf{m}_0, \mathsf{upd}, \mathsf{nxt})$ such that against any Min strategy $\pi \in \Pi_{\mathcal{G}[Y]}$ and from any start state $s \in Y$,

$$\mathcal{P}^{\mathcal{G}[Y]}_{\sigma^*_Y[\mathsf{m}_0],\pi,s}(\mathtt{MP} > 0 \cap \mathtt{EPAR}) = 1$$

3. A uniform positive attractor Max strategy $\sigma_{\mathtt{Attr}}$ that is MD defined for all Max states in X such that against any Min strategy π and any state $s \in X$,

$$\mathcal{P}^{\mathcal{G}}_{\sigma_{\mathtt{Attr}},\pi,s}(\mathsf{F}(S(d))) > 0$$

For small fixed probabilities $\varepsilon_0, \varepsilon_1 > 0$, we define a parameterized family of finite-memory Max strategies $\sigma^*_{\varepsilon_0,\varepsilon_1} \stackrel{\text{def}}{=} (\mathsf{m}_S \uplus \mathsf{M}, \mathsf{upd}', \mathsf{nxt}')$ where

$$\mathsf{nxt}'(\mathsf{m}, s) \stackrel{\text{def}}{=} \begin{cases} \varepsilon_0 \cdot \sigma_{\mathtt{Attr}}(s) + (1-\varepsilon_0) \cdot \sigma_{\mathtt{MP}}(s) & s \in X \\ \sigma_{\mathtt{MP}}(s) & s \in Y,\ \mathsf{m} = \mathsf{m}_S \\ \mathsf{nxt}(\mathsf{m}, s) & s \in Y,\ \mathsf{m} \in \mathsf{M} \end{cases}$$

$$\mathsf{upd}'(\mathsf{m}, (s, s')) \stackrel{\text{def}}{=} \begin{cases} \mathsf{upd}(\mathsf{m}, (s, s')) & \mathsf{m} \in \mathsf{M},\ s \text{ and } s' \in Y \\ \mathsf{m}_S & \mathsf{m} \in \mathsf{M},\ s \text{ or } s' \in X \\ \varepsilon_1 \cdot \mathsf{m}_0 + (1-\varepsilon_1) \cdot \mathsf{m}_S & \mathsf{m} = \mathsf{m}_S,\ s' \in Y \\ \mathsf{m}_S & \mathsf{m} = \mathsf{m}_S,\ s' \in X \end{cases}$$

If $S_\Box \cap X$ is empty, ε_0 is redundant. If $S_\Box \cap Y$ is empty, M and ε_1 are redundant. Observe that when there is only one even color, then indeed Y must be empty, because otherwise Max cannot win almost surely from Y. In this case, $\text{bits}(\sigma^*_{\varepsilon_0,\varepsilon_1})$ is determined by how small ε_0 has to be. We show below that exponentially small ε_0 suffices. For ε_1, we assume that #(distinct even colors) is at least 2. We argue that there is an instantiation with sufficiently small (doubly exponentially small numbers suffice as we will see) positive values for ε_1 such that $\sigma^*_{\varepsilon_0,\varepsilon_1}$ is almost surely winning from any start state. Specifically, we show that it is possible to instantiate ε_0 and ε_1 such that

$$\|\varepsilon_0\| = f_2(n, p_0) \tag{5.7}$$

$$\|\varepsilon_1\| = \mathcal{O}(\|\mathcal{G}\|^{c_0} + \text{bits}(\sigma^*_Y)) \tag{5.8}$$

where f_2 is some polynomial function in two variables, n is the number of states in $\mathcal{G}$, p_0 is the bit size of the probability transition function in $\mathcal{G}$ and c_0 is some constant independent of the instance. From the definition of σ^*, one has $\|\sigma^*\| = \|\sigma^*_Y\| + 1$.

$$\begin{aligned}\text{bits}(\mathsf{nxt}') &= (\|\varepsilon_0\| + \|1-\varepsilon_0\|)|X|(\|\sigma^*_Y\| + 1) + \text{bits}(\mathsf{nxt})\\ \text{bits}(\mathsf{upd}') &= \text{bits}(\mathsf{upd}) + (\|\varepsilon_1\| + \|1-\varepsilon_1\|)|Y|\end{aligned}$$

Substituting for $\|\varepsilon_0\|$ and $\|\varepsilon_1\|$ from above, we get

$$\begin{aligned}\text{bits}(\sigma^*) &= \mathcal{O}(f_2(n,p_0)\|\sigma^*_Y\||S| + \text{bits}(\sigma^*_Y) + (\|\mathcal{G}\|^{c_0} + \text{bits}(\sigma^*_Y))|S|)\\ &= \mathcal{O}(\|\mathcal{G}\|^{c_1}\|\sigma^*_Y\| + \text{bits}(\sigma^*_Y)\|\mathcal{G}\|)\end{aligned}$$

Both size and bit complexity of σ^* satisfy the properties from Item 2.. The proof of claim is complete once we show the required bounds and that σ^* is almost surely winning $\texttt{MP} > 0 \cap \texttt{EPAR}$ from every state. Observe that

$$\forall\pi\, \mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}(\texttt{MP} > 0 \cap \texttt{EPAR}) = 1 \iff \forall\pi \mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}(\texttt{EPAR}) = 1 \wedge \mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}(\texttt{MP} > 0) = 1.$$

For $\texttt{EPAR}$, it suffices to have $\varepsilon_0 > 0$ since

$$\mathcal{P}^{\mathcal{G}}_{\sigma^*,\pi,s}(\texttt{EPAR}) = \mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}(\texttt{EPAR} \cap \mathsf{G}\,\mathsf{F}\,X) + \mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}\big(\texttt{EPAR} \cap \overline{\mathsf{G}\,\mathsf{F}\,X}\big)$$

$$
\begin{aligned}
&= \mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}(\texttt{EPAR} \cap \mathsf{G}\,\mathsf{F}\,X) && +\mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}\big(\overline{\mathsf{G}\,\mathsf{F}\,X}\big) && \sigma^* \text{ eventually behaves as } \sigma^*_Y \\
&= \mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}(\texttt{EPAR} \cap \mathsf{G}\,\mathsf{F}\,S(d)) && +\mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}\big(\overline{\mathsf{G}\,\mathsf{F}\,X}\big) && \textit{Claim } 5.7 \textit{ Item } 1. \text{ as } \varepsilon_0 > 0 \\
&= \mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}(\mathsf{G}\,\mathsf{F}\,S(d)) && +\mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}\big(\overline{\mathsf{G}\,\mathsf{F}\,X}\big) && \\
&= \mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}(\mathsf{G}\,\mathsf{F}\,X) && +\mathcal{P}^{\mathcal{G}}_{\sigma^*[\mathsf{m}_0],\pi,s}\big(\overline{\mathsf{G}\,\mathsf{F}\,X}\big) && \textit{Claim } 5.7 \\
&= 1
\end{aligned}
$$

For $\texttt{MP} > 0$, our argument uses the fact that the strategy σ^* has finite memory and the size of the probabilities it uses. Firstly, observe that $\mathcal{G}[Y]$ is a mean-payoff-parity game with highest color $< d$. Now, to show that $\mathcal{P}^{\mathcal{G}}_{\sigma^*,\pi,s}(\texttt{MP} > 0) = 1$ against any strategy π of Min, it suffices to consider the induced minimizing MDP $\mathcal{M} \stackrel{\text{def}}{=} \mathcal{G}^{\sigma^*}$ and show that the value of every state (s, m_0) in $\mathcal{M}$ is 1. Since the original game has finitely many states and σ^* has finite memory, $\mathcal{M}$ is a finite-state MDP. By standard results on finite-state MDPs for $\texttt{MP} > 0$ [BBE10a], it suffices to consider just MD strategies for Min in this MDP, resulting in a finite-state Markov chain.

Fix some MD strategy $\pi_{\mathcal{M}}$ of Min in $\mathcal{M}$ resulting in the Markov chain $\mathcal{A}[\sigma^*, \pi_{\mathcal{M}}]$ (henceforth referred to as $\mathcal{A}$). Given a state $s \in S$, let $\mathcal{B}$ be some reachable BSCC of $\mathcal{A}$ when starting from (s, m_0). From the definition of σ^*, the memory part of the state is m_S whenever the state belongs to X, except possibly at the beginning. This implies that in $\mathcal{A}$, every state in $X \times \mathsf{M}$ is a transient state, hence cannot be part of $\mathcal{B}$. $\mathcal{B}$ can therefore be seen as disjoint union of $B_1 \stackrel{\text{def}}{=} S_{\mathcal{B}} \cap S \times \mathsf{m}_S$ and $B_2 \stackrel{\text{def}}{=} S_{\mathcal{B}} \cap Y \times \mathsf{M}$. While B_1 and B_2 are disjoint, there will nevertheless be transitions from B_1 to B_2 and vice-versa, if both are non-empty as $\mathcal{B}$ is strongly connected.

We show that $\texttt{MP} > 0$ by a case by case basis on whether either of B_i is empty. If B_2 is empty, this implies that the 'memory part' of the state is m_S and here σ^* behaves as $\sigma_{\texttt{MP}}$ with sufficiently high probability $(1 - \varepsilon_0)$ so that $\texttt{MP} > 0$ almost surely. From this one can estimate the size of ε_0.

▷ Claim 5.11. There are polynomial functions in two variables $f_1(x, y)$, $f_2(x, y)$ such that

- $\|\mu\| \leq f_1(n, p_0)$
- $\|\varepsilon_0\| \leq f_2(n, p_0)$

This shows (5.7)

Proof. When Max plays $\sigma_{\mathtt{MP}}$ solely, the mean-payoff μ achieved against any pure strategy of Min cannot be arbitrarily small. Because the size of the resulting Markov chain is same as the size of the game, and the minimum possible mean-payoff in any BSCC of this Markov chain can be computed through a linear program, it follows that there is a polynomial function f_1 such that $\|\mu\| \leq f_1(n, p_0)$ where p_0 denotes the size of probabilities in $\mathcal{G}$ and n the number of states. But this also means that $\|\varepsilon_0\|$ which is polynomial in $\|\mathcal{G}\|$ suffices since small perturbations of a Markov chain lead only to small changes in the value of achievable mean-payoff. ◁

On the other hand, when B_1 is empty, σ^* behaves as σ^*_Y and B_1 being empty implies $\pi_{\mathcal{M}}$ is also a valid strategy in $\mathcal{G}[Y]$. By properties of σ^*_Y, this means $\mathtt{MP} > 0$ almost surely.

If both B_1 and B_2 are non-empty, then we argue by deriving lower bounds on the expected sum of rewards in B_1 and B_2 under steady state distribution. Recall that $X^s_{i,\mathcal{B}}$ denote the state at time i and $Y^s_{i,\mathcal{B}}$ denote the random variable which computes the sum of the rewards until step i when starting from $s \in \mathcal{B}$. Similarly, $T^s_{2,\mathcal{B}}$ (resp. $T^q_{1,\mathcal{B}}$) denote the hitting times of B_2 (resp. B_1) when starting from $s \in B_1$ (resp. $q \in B_2$). Note that although the states of $\mathcal{B}$ are tuples, we refer to them by s and q for notational simplicity.

$$Y^s_{i,\mathcal{B}} \stackrel{\text{def}}{=} \sum_{j=0}^{i-1} r\big(\big(X^s_{j,\mathcal{B}}, X^s_{j+1,\mathcal{B}}\big)\big) \qquad \text{for all } s \in \mathcal{B},\, i \geq 0 \tag{5.9}$$

$$T^s_{2,\mathcal{B}} \stackrel{\text{def}}{=} \min\big\{i \mid X^s_{i,\mathcal{B}} \in B_2\big\} \qquad s \in B_1 \tag{5.10}$$

$$T^q_{1,\mathcal{B}} \stackrel{\text{def}}{=} \min\big\{i \mid X^q_{i,\mathcal{B}} \in B_1\big\} \qquad q \in B_2 \tag{5.11}$$

If one can find uniform constants $v_1 > 0$ and v_2 such that

$$\forall\, s \quad \mathcal{E}^{\mathcal{B}}_s\Big(Y^s_{T^s_{2,\mathcal{B}},\mathcal{B}}\Big) \geq v_1 \tag{5.12}$$

$$\forall\, q \quad \mathcal{E}^{\mathcal{B}}_q\Big(Y^q_{T^q_{1,\mathcal{B}},\mathcal{B}}\Big) \geq v_2 \tag{5.13}$$

$$v_1 + v_2 > 0 \tag{5.14}$$

then it suffices since MP almost surely equals $\mathcal{E}(\mathtt{MP})$ within a BSCC and

$$\begin{aligned}\mathcal{E}^{\mathcal{B}}(\mathtt{MP}) &= \frac{\sum_s \lambda_s Y^s_{T^s_{2,\mathcal{B}},\mathcal{B}} + \sum_q \nu_q Y^q_{T^q_{1,\mathcal{B}},\mathcal{B}}}{\sum_s \lambda_s T^s_{2,\mathcal{B}} + \sum_q \nu_q T^q_{1,\mathcal{B}}} \\ &\geq \frac{\sum_s \lambda_s v_1 + \sum_q \nu_q v_2}{\sum_s \lambda_s T^s_{2,\mathcal{B}} + \sum_q \nu_q T^q_{1,\mathcal{B}}} \\ &\geq \frac{v_1 + v_2}{\sum_s \lambda_s T^s_{2,\mathcal{B}} + \sum_q \nu_q T^q_{1,\mathcal{B}}} \\ &> 0\end{aligned}$$

where λ_s (resp. ν_q) are long term steady state probabilities of hitting B_1 (resp. B_2) at s (resp. q).

We have $\mathcal{E}^{\mathcal{B}}_q\Big(Y^q_{T^q_{1,\mathcal{B}},\mathcal{B}}\Big) \geq -R\mathcal{E}^{\mathcal{B}}_q\big(T^q_{1,\mathcal{B}}\big)$. To upper bound $T^q_{1,\mathcal{B}}$, observe that $|B_2| \leq |Y| \times |\mathsf{M}|$. Let p_Y denote the maximum size of any probability used by σ^*_Y and x_Y denote the smallest probability in $\mathcal{A}$ restricted to $Y \times \mathsf{M}$. It is easy to see that $x_Y \geq 2^{-p_Y}$ and

$$\mathrm{bits}(\sigma^*_Y) = \mathcal{O}(|Y|\|\sigma^*_Y\|p_Y). \tag{5.15}$$

The probability that a state with transition to B_1 is hit in the first $|B_2|$ steps is at least ${x_Y}^{|B_2|}$ from any start state, *i.e.*, continuing for every b steps of that size we have

$$\forall q \in B_2\ \mathcal{P}^{\mathcal{B}}_q\big(T^q_{1,\mathcal{B}} \geq b \cdot |B_2|\big) \leq \big(1 - {x_Y}^{|B_2|}\big)^b$$

This gives $\mathcal{E}^{\mathcal{B}}_q\big(T^q_{1,\mathcal{B}}\big) \leq {x_Y}^{-|B_2|} \cdot |B_2|$ Therefore

$$\forall q \in B_2\ \mathcal{E}^{\mathcal{B}}_q\Big(Y^q_{T^q_{1,\mathcal{B}},\mathcal{B}}\Big) \geq -{x_Y}^{-|B_2|} \cdot |B_2| \cdot R \tag{5.16}$$

To find a lower bound for $\mathcal{E}^{\mathcal{B}}_s\Big(Y^s_{T^s_{2,\mathcal{B}},\mathcal{B}}\Big)$, we need to link the reward gained in B_1 in the Markov chain $\mathcal{A}$ to reward gained in B_1 in the Markov chain $\mathcal{A}' \stackrel{\text{def}}{=} \mathcal{G}[\sigma_1, \pi']$ where σ_1 and π' are memoryless restrictions of σ^* and $\pi_{\mathcal{M}}$ to $S \times \mathsf{m}_S$. Also note that $\mathcal{A}'[B_1]$ is a well-defined sub-Markov chain and is equivalent to the conditioned Markov chain of $\mathcal{A}$ on B_1.

▶ **Definition 5.12** (Conditioned Markov chain)**.** *If $\mathcal{A} \stackrel{def}{=} (S, E, P)$ is some finite-state Markov chain and $S' \subseteq S$ such that for all $s' \in S'$, $\sum_{s \in S'} P((s', s)) > 0$, then the conditioned Markov chain is well defined and given by $\mathcal{A}_{S'} \stackrel{def}{=} (S', E \cap (S' \times S'), P_{S'})$, where $P_{S'}((s', s)) \stackrel{def}{=} \frac{P(s',s)}{\sum_{s \in S'} P((s',s))}$*

Denote by $r^{\mathcal{B}}_\ell \stackrel{\text{def}}{=} r\big(\big(X^s_{l,\mathcal{B}}, X^s_{l+1,\mathcal{B}}\big)\big)$ the random variable which gives the reward on the ℓ^{th} step when starting from $s \in \mathcal{B}$. For an event A, let $\mathbf{1}_A$ denote the

indicator random variable of A. Simplifying, we get $\mathcal{E}_s^{\mathcal{B}}\Big(Y^s_{T^s_{2,\mathcal{B}},\mathcal{B}}\Big)$

$$
\begin{aligned}
&= \mathcal{E}_s^{\mathcal{B}}\left(\sum_{\ell=0}^{\infty} r_\ell^{\mathcal{B}} \cdot \mathbf{1}_{T^s_{2,\mathcal{B}}>\ell}\right) && \text{from (5.9)}\\
&= \sum_{\ell=0}^{\infty} \mathcal{E}_s^{\mathcal{B}}\Big(r_\ell^{\mathcal{B}} \cdot \mathbf{1}_{T^s_{2,\mathcal{B}}>\ell}\Big) && |r_\ell^{\mathcal{B}}| \le R\ ,\ \mathcal{E}_s^{\mathcal{B}}\big(T^s_{2,\mathcal{B}}\big) < \infty\\
&= \sum_{\ell=0}^{\infty} \mathcal{E}_s^{\mathcal{B}}\Big(r_\ell^{\mathcal{B}} \cdot \mathbf{1}_{T^s_{2,\mathcal{B}}>\ell+1}\Big) + \mathcal{E}_s^{\mathcal{B}}\Big(r_\ell^{\mathcal{B}} \cdot \mathbf{1}_{T^s_{2,\mathcal{B}}=\ell+1}\Big) && T^s_{2,\mathcal{B}} \text{ is integer valued}
\end{aligned}
$$

The value $\mathcal{E}_s^{\mathcal{B}}\Big(r_\ell^{\mathcal{B}} \cdot \mathbf{1}_{T^s_{2,\mathcal{B}}=\ell+1}\Big)$ can be lower bounded by $-R \cdot \mathcal{P}_s^{\mathcal{B}}\big(T^s_{2,\mathcal{B}} = \ell+1\big)$ since $r_\ell^{\mathcal{B}} \ge -R$ surely. Furthermore, it is easy to see from (5.10) that the event $T^s_{2,\mathcal{B}} > \ell+1$ is exactly $\bigcap_{j=0}^{\ell+1} X^s_{j,\mathcal{B}} \in B_1$. Denoting the latter event by $\mathsf{G}^{[0,\ell+1]}B_1$ and continuing,

$$
\begin{aligned}
&\ge \sum_{\ell=0}^{\infty}\big(\mathcal{E}_s^{\mathcal{B}}\big(r_\ell^{\mathcal{B}} \cdot \mathbf{1}_{\mathsf{G}^{[0,\ell+1]}B_1}\big) - R \cdot \mathcal{P}_s^{\mathcal{B}}\big(T^s_{2,\mathcal{B}} = \ell+1\big)\big)\\
&\ge \left(\sum_{\ell=0}^{\infty} \mathcal{E}_s^{\mathcal{B}}\big(r_\ell^{\mathcal{B}} \cdot \mathbf{1}_{\mathsf{G}^{[0,\ell+1]}B_1}\big)\right) - R && s \in B_1
\end{aligned}
$$

Intuitively, the expectation of the sum as long as one stays in B_1 should be the same as the expectation of the sum in the conditioned Markov chain $\mathcal{A}_{B_1}$ (this is also the reason to throw away the last transition). We formalize this notion.

▷ Claim 5.13. For all $\ell \ge 0$, $s \in B_1$

$$\mathcal{E}_s^{\mathcal{B}}\big(r_\ell^{\mathcal{B}} \cdot \mathbf{1}_{\mathsf{G}^{[0,\ell+1]}B_1}\big) = \mathcal{E}_s^{\mathcal{A}_{B_1}}\Big(r_\ell^{\mathcal{A}_{B_1}}\Big) \cdot \mathcal{P}_s^{\mathcal{B}}\big(\mathsf{G}^{[0,\ell+1]}B_1\big)$$

Proof of Claim 5.13. For succinctness let $E_{\ell+1} = \mathsf{G}^{[0,\ell+1]}B_1$. The claim follows if one proves that for any $s_1, s_2 \in B_1$

$$\mathcal{P}_s^{\mathcal{B}}\big(X^s_{\ell,\mathcal{B}} = s_1, X^s_{\ell+1,\mathcal{B}} = s_2 \;\big|\; E_{\ell+1}\big) = \mathcal{P}_s^{\mathcal{A}_{B_1}}\Big(X^s_{\ell,\mathcal{A}_{B_1}} = s_1, X^s_{\ell+1,\mathcal{A}_{B_1}} = s_2\Big)$$

holds. It is easy to see why as the rewards on the identical transitions are equal in both Markov chains. To show above, we first start by showing that the distribution of states at the ℓ^{th} step is identical in both scenarios. Given ℓ and start state s, for all $s_1 \in B_1$ let $\theta_{s,\ell}(s_1) \stackrel{\text{def}}{=} \mathcal{P}_s^{\mathcal{B}}\big(X^s_{\ell,\mathcal{B}} = s_1 \;\big|\; E_{\ell+1}\big)$ and $\theta'_{s,\ell}(s_1) \stackrel{\text{def}}{=} \mathcal{P}_s^{\mathcal{A}_{B_1}}\Big(X^s_{\ell,\mathcal{A}_{B_1}} = s_1\Big)$.

▷ Claim 5.14. For all $\ell \ge 0$, $s \in B_1$

$$\theta_{s,\ell} = \theta'_{s,\ell}$$

Proof of Claim 5.14. We prove by induction on ℓ.

Base case : When $\ell = 0$, the event $X^s_{1,\mathcal{B}} \in B_1$ has positive probability by properties of B_1 and therefore $\theta_{s,\ell}$ is well-defined for every start state $s \in B_1$ and is equal to δ_s. This proves the base case.

Induction step : Assume, $\theta_{s,\ell} = \theta'_{s,\ell}$ for some $\ell \geq 0$. To show that it holds for $\ell + 1$, $\theta_{s,\ell+1}(s_2)$

$$
\begin{aligned}
&= \mathcal{P}^{\mathcal{B}}_s\big(X^s_{\ell+1,\mathcal{B}} = s_2 \;\big|\; E_{\ell+1}\big) \\
&= \sum_{s_1 \in B_1} \mathcal{P}^{\mathcal{B}}_s\big(X^s_{\ell+1,\mathcal{B}} = s_2 \;\big|\; E_{\ell+1}, X^s_{\ell,\mathcal{B}} = s_1\big) \cdot \theta_{s,\ell}(s_1) \\
&= \sum_{s_1 \in B_1} \mathcal{P}^{\mathcal{B}}_{s_1}\big(X^{s_1}_{1,\mathcal{B}} = s_2 \;\big|\; X^{s_1}_{1,\mathcal{B}} \in B_1\big) \cdot \theta'_{s,\ell}(s_1) && \text{Markov, IH} \\
&= \sum_{s_1 \in B_1} \frac{P_{\mathcal{B}}((s_1, s_2))}{\sum_{s_3 \in B_1} P_{\mathcal{B}}((s_1, s_3))} \cdot \theta'_{s,\ell}(s_1) \\
&= \sum_{s_1 \in B_1} P_{B_1}((s_1, s_2)) \cdot \theta'_{s,\ell}(s_1) && \textit{Definition } 5.12 \\
&= \theta'_{s,\ell+1}(s_2)
\end{aligned}
$$

◁

Now, $\mathcal{P}^{\mathcal{B}}_s\big(X^s_{\ell,\mathcal{B}} = s_1, X^s_{\ell+1,\mathcal{B}} = s_2 \;\big|\; E_{\ell+1}\big)$

$$
\begin{aligned}
&= \mathcal{P}^{\mathcal{B}}_s\big(X^s_{\ell+1,\mathcal{B}} = s_2 \;\big|\; E_{\ell+1}, X^s_{\ell,\mathcal{B}} = s_1\big) \cdot \theta_{s,\ell}(s_1) \\
&= P_{B_1}((s_1, s_2)) \cdot \theta'_{s,\ell}(s_1) && \textit{Claim } 5.14 \\
&= \mathcal{P}^{\mathcal{A}_{B_1}}_s\Big(X^s_{\ell,\mathcal{A}_{B_1}} = s_1, X^s_{\ell+1,\mathcal{A}_{B_1}} = s_2\Big)
\end{aligned}
$$

Therefore, one has $\mathcal{E}^{\mathcal{B}}_s\big(r^{\mathcal{B}}_\ell \;\big|\; E_{\ell+1}\big) = \mathcal{E}^{\mathcal{A}_{B_1}}_s\Big(r^{\mathcal{A}_{B_1}}_\ell\Big)$ from which the claim follows.

◁

From Claim 5.13, we further get $\mathcal{E}_s^{\mathcal{B}}\left(Y^s_{T^s_{2,\mathcal{B}},\mathcal{B}}\right)$

$$
\begin{aligned}
&\geq \left(\sum_{\ell=0}^{\infty} \mathcal{E}_s^{\mathcal{A}_{B_1}}\left(r_\ell^{\mathcal{A}_{B_1}}\right) \cdot \mathcal{P}_s^{\mathcal{B}}\left(T^s_{2,\mathcal{B}} > \ell+1\right)\right) - R \\
&\geq \left(\sum_{\ell=0}^{\infty} \mathcal{E}_s^{\mathcal{A}_{B_1}}\left(r_\ell^{\mathcal{A}_{B_1}}\right) \cdot \sum_{t=\ell+2}^{\infty} \mathcal{P}_s^{\mathcal{B}}\left(T^s_{2,\mathcal{B}} = t\right)\right) - R \\
&\geq \left(\sum_{t=2}^{\infty} \mathcal{P}_s^{\mathcal{B}}\left(T^s_{2,\mathcal{B}} = t\right) \cdot \left(\sum_{\ell=0}^{t-2} \mathcal{E}_s^{\mathcal{A}_{B_1}}\left(r_\ell^{\mathcal{A}_{B_1}}\right)\right)\right) - R && \text{Interchange sums} \\
&\geq \left(\sum_{t=2}^{\infty} \mathcal{P}_s^{\mathcal{B}}\left(T^s_{2,\mathcal{B}} = t\right) \cdot \mathcal{E}_s^{\mathcal{A}_{B_1}}\left(Y^s_{t-1,\mathcal{A}_{B_1}}\right)\right) - R && \text{from (5.9) and linearity} \\
&\geq \left(\sum_{t=1}^{\infty} \mathcal{P}_s^{\mathcal{B}}\left(T^s_{2,\mathcal{B}} = t\right) \cdot \mathcal{E}_s^{\mathcal{A}_{B_1}}\left(Y^s_{t-1,\mathcal{A}_{B_1}}\right)\right) - R && Y^s_{0,\mathcal{A}_{B_1}} = 0
\end{aligned}
$$

This allows us to turn a lower bound on sums in $\mathcal{A}_{B_1}$ into a bound for $\mathcal{E}_s^{\mathcal{B}}\left(Y^s_{T^s_{2,\mathcal{B}},\mathcal{B}}\right)$. Remember that $\mathcal{A}_{B_1}$ is a sub-Markov chain of the Markov chain $\mathcal{A}' = \mathcal{G}[\sigma_1, \pi']$ induced by two memoryless strategies. This means, the size of the probabilities in $\mathcal{A}'$ is $p_1 \stackrel{\text{def}}{=} \max(\|\varepsilon_0\|, p_0) = \|\varepsilon_0\|$ and let the smallest probability be $x_1 \stackrel{\text{def}}{=} \varepsilon_0$. A lower bound on sums in $\mathcal{A}'$ is also a lower bound for sums in $\mathcal{A}_{B_1}$. Let $\mathcal{A}' = (S, E', P')$. Fix a start state s and denote by μ', the minimum achievable mean payoff in any BSCC in $\mathcal{A}'$. By optimality of σ_1, $\mu' > 0$. We can split the sum into two parts: the sum within a BSCC and the sum in transient states. For the former, within any BSCC B of $\mathcal{A}'$, one can get a lower bound by standard arguments using the martingale process obtained from the Poisson equation Lemma 2.7 and Fact 1.

▶ **Definition 5.15.** *For a function $f \in \mathbb{R}^S$ on an ergodic Markov chain $\mathcal{A} = (S, E, P)$, the Poisson equation is given by*

$$\boldsymbol{f} + P\boldsymbol{h} = \boldsymbol{h} + \mathbf{1}\bar{f}$$

where $\bar{f} = \lim_{n\to\infty} \frac{\sum_{i=0}^{n-1} f\left(X_{i,\mathcal{A}}\right)}{n}$ is the long term average of f.

By standard results(Lemma 2.7), a solution for the above equation exists and $\boldsymbol{h}$ can be chosen such that $h(s) \in [0, K]$ where $K = \frac{2|S| f_{\max}}{x_0^{|S|}}$, $f_{\max} = \max_s |f(s)|$, x_0 the minimum non-zero probability in P. Moreover, given such a h, the process $M_n = \sum_{i=0}^{n-1} f(X_{i,\mathcal{A}}) + h(X_{n,\mathcal{A}}) - n\bar{f}$ is a martingale.

To use these results in our context, lets fix a BSCC B and a start state s_B. Within this BSCC, let the long term average mean payoff be $\mu \geq \mu' > 0$.

Denote by $h^B_{\max} \stackrel{\text{def}}{=} \frac{2|S_B|R^B}{p_0^{B|S_B|}}$. Then one can find $h^B : S_B \to \left[0, h^B_{\max}\right]$ such that $M^B_t = Y^{s_B}_{t,B} + h^B\left(X^{s_B}_{t,B}\right) - t\mu$ is a martingale.

$$\implies \mathcal{E}^B_{s_B}\left(Y^{s_B}_{t,B} + h^B\left(X^{s_B}_{t,B}\right) - t\mu\right) = \mathcal{E}^B_{s_B}\left(M^B_t\right) = \mathcal{E}^B_{s_B}\left(M^B_0\right) = h^B(s_B)$$

Simplifying, we get

$$\forall\, s_B,\, \mathcal{E}^B_{s_B}\left(Y^{s_B}_{t,B}\right) = t\mu + h^B(s_B) - \mathcal{E}^B_{s_B}\left(h^B\left(X^{s_B}_{t,B}\right)\right) \geq t\mu' - h^B_{\max}$$

Denote by H, the set of all states which are part of some BSCC in $\mathcal{A}'$ and $T^s_H \stackrel{\text{def}}{=} \min\left\{i \geq 0 \mid X^s_{i,\mathcal{A}'} \in H\right\}$ denote the hitting time to one of these BSCC's. $|S| = n$, let $h^{\mathcal{A}'}_{\max} \stackrel{\text{def}}{=} \max_{B \text{ is a BSCC}} h^B_{\max} \leq \frac{2nR}{x_1^n}$, then it is clear that for any $s \in H$

$$\mathcal{E}^{\mathcal{A}'}_s\left(Y^s_{t,\mathcal{A}'}\right) \geq t\mu' - h^{\mathcal{A}'}_{\max} \tag{5.17}$$

and define $C_{\mathcal{A}'} \stackrel{\text{def}}{=} (R + \mu')\frac{n}{(x_1)^n} + \frac{2nR}{(x_1)^n} = \frac{3nR + n\mu'}{(x_1)^n}$. (5.17) provides a lower bound for all the states in H. To get a lower bound on the sum for the transient states, we first compute an upper bound on the expected time spent in these states with the analysis similar to Lemma 2.15.

▷ Claim 5.16. For any state s in $\mathcal{A}'$,

$$\mathcal{E}^{\mathcal{A}'}_s(T^s_H) \leq \frac{n}{x_1^n}$$

Proof. Much like in the derivation for (5.16), one can show that $\mathcal{P}^{\mathcal{A}'}_s(T^s_H > k \cdot n) \leq (1 - x_1^n)^k$. This then implies

$$\begin{aligned}\mathcal{E}^{\mathcal{A}'}_s(T^s_H) &= \sum_{k=0}^{\infty} \mathcal{P}^{\mathcal{A}'}_s(T^s_H > k) \leq n \sum_{k=0}^{\infty} \mathcal{P}^{\mathcal{A}'}_s(T^s_H > k \cdot n) \\ &= n \sum_{k=0}^{\infty} (1 - x_1^n)^k = \frac{n}{x_1^n}\end{aligned}$$

◁

▷ Claim 5.17.

$$\mathcal{E}^{\mathcal{A}'}_s\left(Y^s_{t,\mathcal{A}'}\right) \geq t\mu' - C_{\mathcal{A}'}$$

Proof of Claim 5.17. We decompose the expected reward $\mathcal{E}^{\mathcal{A}'}_s\left(Y^s_{t,\mathcal{A}'}\right)$ by conditioning on T^s_H. For a given $T^s_H = k$, the reward is a sum of rewards from the transient phase (at least $-R \cdot k$) and the recurrent phase. For the recurrent

phase, the expected reward in the recurrent phase of length $t-k$ is at least $(t-k)\mu' - h^{\mathcal{A}'}_{\max}$ (Equation (5.17)).

$$\begin{aligned}
\mathcal{E}^{\mathcal{A}'}_s\big(Y^s_{t,\mathcal{A}'}\big) &= \mathcal{E}^{\mathcal{A}'}_s\Big(\mathcal{E}^{\mathcal{A}'}_s\big(Y^s_{t,\mathcal{A}'} \mid T^s_H\big)\Big) \\
&\geq \mathcal{E}^{\mathcal{A}'}_s\Big(-R\cdot T^s_H + (t-T^s_H)\mu' - h^{\mathcal{A}'}_{\max}\Big) \\
&= t\mu' - (R+\mu')\mathcal{E}^{\mathcal{A}'}_s(T^s_H) - h^{\mathcal{A}'}_{\max} \\
&\geq t\mu' - (R+\mu')\frac{n}{x_1^n} - h^{\mathcal{A}'}_{\max}\ (Claim\ 5.16) \\
&\geq t\mu' - C_{\mathcal{A}'}
\end{aligned}$$

◁

Using Claim 5.17, $\mathcal{E}^{\mathcal{B}}_s\Big(Y^s_{T^s_{2,\mathcal{B}},\mathcal{B}}\Big)$

$$\begin{aligned}
&\geq \left(\sum_{t=1}^{\infty}\mathcal{P}^{\mathcal{B}}_s\big(T^s_{2,\mathcal{B}}=t\big)\cdot(t-1)\mu' - C_{\mathcal{A}'}\right) - R \\
&\geq \mu'\cdot\left(\sum_{t=1}^{\infty}\mathcal{P}^{\mathcal{B}}_s\big(T^s_{2,\mathcal{B}}=t\big)\cdot t\right) - C_{\mathcal{A}'} - \mu' - R \qquad\qquad s\in B_1 \\
&= \mu'\cdot\mathcal{E}^{\mathcal{B}}_s\big(T^s_{2,\mathcal{B}}\big) - C_{\mathcal{A}'} - \mu' - R
\end{aligned}$$

It is easy to see that $\mathcal{E}^{\mathcal{B}}_s\big(T^s_{2,\mathcal{B}}\big) \geq \frac{1}{\varepsilon_1}$. Substituting this, we get

$$\mathcal{E}^{\mathcal{B}}_s\Big(Y^s_{T^s_{2,\mathcal{B}},\mathcal{B}}\Big) \geq \frac{\mu'}{\varepsilon_1} - C_{\mathcal{A}'} - \mu' - R \tag{5.18}$$

Comparing (5.18) with (5.12), and (5.16) with (5.13), and looking at (5.14), we require ε_1 to be such that

$$\frac{\mu'}{\varepsilon_1} - C_{\mathcal{A}'} - \mu' - R - {x_Y}^{-|B_2|}\cdot|B_2|\cdot R > 0$$

One can consider ε_1 such that

$$\varepsilon_1 = \frac{\mu'}{\lceil C_{\mathcal{A}'} + \mu' + R + {x_Y}^{-|B_2|}\cdot|B_2|\cdot R\rceil + 1} \tag{5.19}$$

μ' and $C_{\mathcal{A}'}$ constants which arise out of $\mathcal{A}'$ whose size is polynomial in $\mathcal{G}$. This implies that the sizes of both μ' and $C_{\mathcal{A}'}$ is bounded by some polynomial with large enough degree. Also, $|B_2| \leq |Y| \times \|\sigma^*_Y\|$. Combining all the above facts, it is easy to see that $\|\varepsilon_1\| = \mathcal{O}(\|\mathcal{G}\|^c + |Y|\cdot\|\sigma^*_Y\|\cdot p_Y)$ for some large enough degree c. From (5.15), this shows that $\|\varepsilon_1\| = \mathcal{O}(\|\mathcal{G}\|^{c_0} + \mathrm{bits}(\sigma^*_Y))$ which is what we sought to prove in (5.8) ◁

In Claim 5.6, $\|\sigma^*\|$ is the maximum of the $\|\sigma_i^*\|$ which, by induction hypothesis, is at most #(distinct even colors). Since, in this case, the maximum color is odd, the number of even colors in the subgames is at most the same as in the whole game. Hence $\|\sigma^*\| = \max_i \|\sigma_i^*\|$. By induction hypothesis $\text{bits}(\sigma_i^*) = \mathcal{O}\Big(\|\mathcal{G}\|^{d-1+c}\Big)$, and from Claim 5.6, we get $\text{bits}(\sigma^*) = \mathcal{O}(|S|\,\text{bits}(\sigma_i^*)) = \mathcal{O}(\|\mathcal{G}\|\,\text{bits}(\sigma_i^*)) = \mathcal{O}(\|\mathcal{G}\|^{d+c})$. This shows Equation (5.3).

When the maximum color d is even and $k = \#(\text{distinct even colors})$, Claim 5.10 implies that the number of even colors in Y is at most $k-1$. Since $\|\sigma^*\| = 1 + \|\sigma_Y^*\|$, we have proven the induction step for the number of memory modes. For the bit size, by the induction hypothesis $\text{bits}(\sigma_Y^*) = \mathcal{O}\big(\|\mathcal{G}\|^{d-1+c}\big)$. From Claim 5.10,

$$\begin{aligned}\text{bits}(\sigma^*) &= \mathcal{O}(\|\mathcal{G}\|^{c_1}\|\sigma_Y^*\| + \text{bits}(\sigma_Y^*)\|\mathcal{G}\|)\\ &= \mathcal{O}\big(\|\mathcal{G}\|^{c_1+1} + \|\mathcal{G}\|^{d+c}\big)\\ &= \mathcal{O}\big(\|\mathcal{G}\|^{d+c}\big)\end{aligned}$$

This shows Equation (5.4). ◀

Towards proving the lower bound, we construct the following class of games (Definition 5.18 and Figure 5.2).

▶ **Definition 5.18.** *For each $n \in \mathbb{Z}_+$ let $Odd^n \stackrel{def}{=} \{x \in \mathbb{Z}_+ \mid x \le 2n \wedge x\%2 = 1\}$ and $Even^n \stackrel{def}{=} \{x \in \mathbb{Z}_+ \mid x \le 2n \wedge x\%2 = 0\}$ and*
$\mathcal{G}_n \stackrel{def}{=} \big(S^n, \big(S^n_\Box, S^n_\Diamond, S^n_\bigcirc\big), E^n, P^n\big)$ where

1. *$S^n_\Box \stackrel{def}{=} \{t\} \uplus Odd^n$, $S^n_\Diamond \stackrel{def}{=} \{s\} \uplus Even^n$, $S^n_\bigcirc \stackrel{def}{=} \emptyset$, consequently P^n is trivial.*
2. *$E^n \stackrel{def}{=} t \times Odd^n \cup Odd^n \times s \cup s \times Even^n \cup Even^n \times t$*

For $n \in \mathbb{Z}_+$, state n has color n and $Col(s) = Col(t) \stackrel{def}{=} 1$. The reward function r^n on the edges is defined as $r^n((t,i)) \stackrel{def}{=} i+1$, $r^n((i,s)) \stackrel{def}{=} 0$, $r^n((s,j)) \stackrel{def}{=} -j+1$, $r^n((j,t)) \stackrel{def}{=} 0$.

Proof of Item 2.(Theorem 5.5). It is clear from Definition 5.18 that $\|\mathcal{G}_n\| = \Theta(n)$ and there are n even colors in $\mathcal{G}_n$ for every $n \ge 2$. Min (resp. Max) controls only one state s (resp. t), where it has a non-trivial choice. Max can win $\texttt{MP} > 0 \cap \texttt{EPAR}$ surely from state s (and thus from every other state) by the strategy that, at state t, copies Min's most recent observed choice at state s. I.e., if Min's last step was $s \to x$ for some even x then Max chooses $t \to (x-1)$.

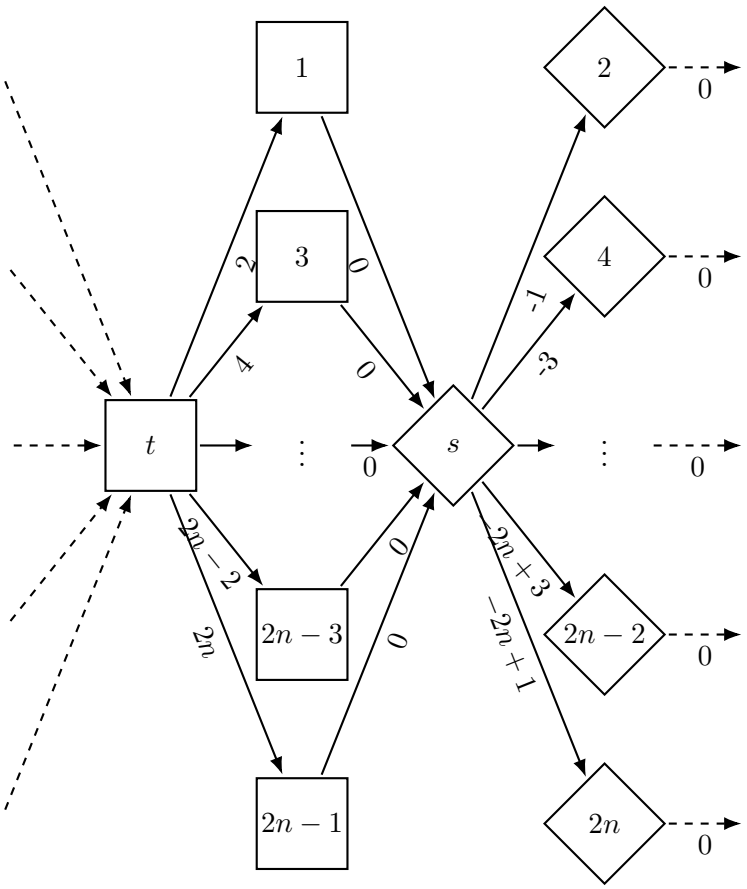

Figure 5.2: Game $\mathcal{G}_n$ from Definition 5.18 with odd states up to $2n-1$ and even states up to $2n$.

(This Max strategy requires n memory modes.) All induced plays trivially satisfy EPAR. Moreover, the total reward between consecutive visits to state s is always $-(x-1)+x=1$, and thus $\mathtt{MP}>0$ is also satisfied surely.

However, we show that all Max strategies with $< n$ memory modes are worthless.

Towards a contradiction, assume that there exists a Max strategy σ from state s in $\mathcal{G}_n$ with a set of memory modes M where $|\mathsf{M}| < n$ and $\inf_\pi \mathcal{P}^{\mathcal{G}_n}_{\sigma,\pi,s}(\mathtt{MP}>0 \cap \mathtt{EPAR}) > 0$.

First we show, by induction on k, the following property $P(k)$ for every $1 \le k \le n-1$: There exists a subset M_k of the memory modes, such that $|\mathsf{M}_k| \ge k$ and for every $\mathsf{m} \in \mathsf{M}_k$, the support of $\sigma[\mathsf{m}](t)$ is limited to the k lowest options, *i.e.*, $\subseteq \{1,3,\ldots,2k-1\}$.

Base case $k=1$: Let π be the Min strategy that always chooses the lowest option 2. If, for every $\mathsf{m} \in \mathsf{M}$, $\sigma[\mathsf{m}](t)$ includes options *different from* 1, then

$\mathcal{P}^{\mathcal{G}_n}_{\sigma,\pi,s}(\texttt{MP} > 0 \cap \texttt{EPAR}) \leq \mathcal{P}^{\mathcal{G}_n}_{\sigma,\pi,s}(\texttt{EPAR}) = 0$, a contradiction. Hence there must exist at least one memory mode $\mathsf{m} \in \mathsf{M}_1$ such that the support of $\sigma[\mathsf{m}](t)$ is limited to option 1.

Induction step $k-1$ to k: Let π be the Min strategy that always chooses the option $2k$. Consider the long-run behavior of the finite-state Markov chain induced by $\mathcal{G}_n$, σ and π. In the runs where Max's memory mode at t is eventually always contained in M_{k-1}, the total payoff between consecutive visits to s is ≤ -1 and therefore $\texttt{MP} < 0$ and thus these runs are losing for $\texttt{MP} > 0 \cap \texttt{EPAR}$. Hence there must exist a non-null set of runs where Max's memory mode at t is infinitely often in $\mathsf{M} \setminus \mathsf{M}_{k-1}$. Suppose that for all $\mathsf{m} \in \mathsf{M} \setminus \mathsf{M}_{k-1}$ the support of $\sigma[\mathsf{m}](t)$ is *not* limited to the k lowest options $\{1, \ldots, 2k-1\}$. Then $\mathcal{P}^{\mathcal{G}_n}_{\sigma,\pi,s}(\texttt{MP} > 0 \cap \texttt{EPAR}) \leq \mathcal{P}^{\mathcal{G}_n}_{\sigma,\pi,s}(\texttt{EPAR}) = 0$, a contradiction. Therefore there exists at least one $\mathsf{m} \in \mathsf{M} \setminus \mathsf{M}_{k-1}$ such that the support of $\sigma[\mathsf{m}](t)$ is limited to the k lowest options $\{1, \ldots, 2k-1\}$. Thus $\mathsf{M}_k \supseteq \{\mathsf{m}\} \uplus \mathsf{M}_{k-1}$ and, by induction hypothesis, $|\mathsf{M}_k| \geq 1 + |\mathsf{M}_{k-1}| \geq 1 + (k-1) = k$. This concludes the induction step.

For $k = n-1$, property $P(n-1)$ yields the required contradiction. Since $|\mathsf{M}| \leq n-1$, we have $\mathsf{M} = \mathsf{M}_{n-1}$. Thus the support of $\sigma(t)$ never includes the highest option $2n-1$. Let π be the Min strategy that always chooses the highest option $2n$. Then $\mathcal{P}^{\mathcal{G}_n}_{\sigma,\pi,s}(\texttt{MP} > 0 \cap \texttt{EPAR}) \leq \mathcal{P}^{\mathcal{G}_n}_{\sigma,\pi,s}(\texttt{MP} > 0) = 0$. ◀

Remark 5.19 (On the Computational Complexity). We did not touch on the computational complexity of finding the almost surely winning set of states. But one can observe that [CDGO14, Algorithm 1] where Mean on line 6 would now mean $\texttt{MP} > 0$, would essentially be an algorithm that finds the required set of states. So the bounds given in the above reference for that algorithm($\mathcal{O}\big(d \cdot n^{2d} \cdot \mathsf{Mean}(n, p_0, R)\big)$) is trivially also an upper bound for $\texttt{MP} > 0 \cap \texttt{EPAR}$.

5.5 Counterexamples for Lifting Randomized Strategies

[GZ09, Theorem 9] describes a different method to lift strategies from MDPs to SSGs. Unlike [GK23, Theorem 6.1], it does not require that the objective is shift-invariant inverse-submixing, but instead has stronger prerequisites about the strategy complexity in MDPs (resp. 1-player games). Given an objective, if both players have optimal memoryless *deterministic* strategies in all maximizing

(resp. minimizing) MDPs, then they also have optimal memoryless deterministic strategies in all SSGs. Under mild assumptions this can be generalized to finite-memory deterministic strategies [BORV23, Theorem 4.1], in stochastic or non-stochastic arenas.

Such results do *not* generalize to *randomized* strategies as discussed in the counterexamples below.

However, first note that there cannot exist any counterexample where the objective is shift-invariant and submixing. In that case [GK23, Theorem 1.1] implies that Max has optimal MD strategies in SSGs. This leaves the question whether there are counterexamples that satisfy other strong properties, e.g., shift-invariant and inverse-submixing.

One counterexample was discussed in [BORV23, Section 4.4]. For the $\texttt{MP} = 0$ objective, Max (resp. Min) have optimal MR (resp. MD) strategies in deterministic 1-player games, but optimal Max strategies require at least 1 bit of memory in deterministic 2-player games, even if randomization is allowed. [BORV23, Section 4.4] does not analyze the strategy complexity of this objective in MDPs, but it is easy to show that Max (resp. Min) has optimal MR (resp. MD) strategies even in MDPs. It is also easy to extend this example with multiple rewards, such that optimal Max strategies require more (but still finite) memory. However, there remains one downside. While the $\texttt{MP} = 0$ objective is shift-invariant, it is neither submixing nor inverse-submixing. (The sequences constructed in the proof of [BBE10b, Lemma 9] provide counterexamples to either property.)

Our lower bounds for the one-dimensional $\texttt{MP} > 0 \cap \texttt{EPAR}$ objective in Section 5.4 show a slightly stronger property. Max (resp. Min) have optimal MR (resp. MD) strategies even in MDPs, but Max strategies in deterministic 2-player games can require arbitrarily many memory modes (equal to the number of even colors), even if randomization is allowed. Moreover, the $\texttt{MP} > 0 \cap \texttt{EPAR}$ objective is shift-invariant and inverse-submixing.

The multi-dimensional $\texttt{MP} > \mathbf{0}$ objective considered in Section 5.3 is also shift-invariant and inverse-submixing and Max (resp. Min) have optimal MR (resp. MD) strategies even in MDPs. However, optimal Max strategies in deterministic 2-player games can require an *exponential* (in the dimension) number of memory modes, even if randomization is allowed.

If one drops the requirement that Min has optimal MD strategies in MDPs, then there exist even stronger counterexamples where Max requires infinite memory

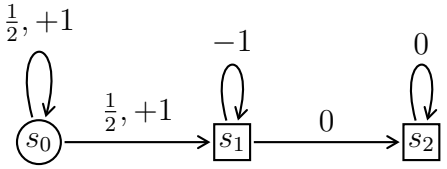

Figure 5.3: An MDP where every optimal Max strategy for the objective W requires infinite memory. In the random state s_0 the successor is either s_0 or s_1, each with probability $1/2$, and the reward is $+1$. In the controlled state s_1 Max chooses between staying in s_1 with reward -1 or going to s_2 with reward 0. State s_2 is a sink with reward 0.

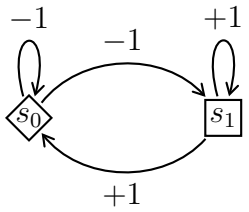

Figure 5.4: A deterministic 2-player game where every optimal Max strategy for the objective $\mathtt{0}$ with $X \stackrel{\text{def}}{=} \{s_0\}$ requires infinite memory. State s_0 (resp. s_1) belongs to Min (resp. Max) with reward -1 (resp. $+1$). The players choose between staying in the current state or switching.

in deterministic 2-player games.

An example in [Van23, Proposition 3.1.3] considers the objective $W \stackrel{\text{def}}{=} W_1 \cup W_2$, where $W_1 \stackrel{\text{def}}{=} \{c_1 c_2 \cdots \mid \liminf_{n\to\infty} \sum_{i=1}^{n} c_i = +\infty\}$ and $W_2 \stackrel{\text{def}}{=} \{c_1 c_2 \cdots \mid \sum_{i=1}^{n} c_i = 0 \text{ for infinitely many } n\}$. Both players have optimal finite-memory deterministic strategies in deterministic 1-player games, but Max needs infinite memory in deterministic 2-player games. However, note that the objective W_2 (and thus $W = W_1 \cup W_2$) is not shift-invariant. Moreover, [Van23] does not discuss the strategy complexity of W in MDPs, and the example in Figure 5.3 shows that optimal Max strategies for W already require infinite memory in MDPs. I.e., W is not a counterexample for lifting strategies from MDPs to SSGs.

▶ **Proposition 5.20.** *Given the MDP in Figure 5.3, Max has an optimal strategy that attains the objective W with probability 1, but every FR Max strategy is not optimal.*

Proof. First, since s_0 is left almost surely, W_1 is a null set under all Max strategies.

An optimal Max strategy plays as follows. When s_1 is reached with total reward $+n$ (which happens with probability 2^{-n}), then it loops n times in s_1 and then goes to s_2 where the total reward stays 0 forever. This satisfies W_2 (and thus W) with probability 1.

Now consider an FR Max strategy with m memory modes. Since m is finite, there exists at least one memory mode m and numbers $n_2 > n_1 > 0$ such that the two events of entering s_1 with memory mode m and total rewards n_1 and n_2 each have a nonzero probability. Thus with nonzero probability either the state s_2 is reached with a total reward $\neq 0$ or s_2 is never reached, and hence W is not satisfied almost surely. ◀

We now present a stronger counterexample where the objective is shift-invariant and both Max and Min each have optimal MR (or alternatively, FDD) strategies in all MDPs, but optimal Max strategies still require infinite memory (even if randomization is allowed) in deterministic 2-player games.

Let $\mathtt{O}_1 \stackrel{\text{def}}{=} \{c_1 c_2 \cdots \mid \limsup_{n\to\infty} \sum_{i=1}^{n} c_i > -\infty\}$. Let $X \subseteq S$ be a subset of the states, e.g., defined via a coloring function. Our objective $\mathtt{O}$ combines $\mathtt{O}_1$ with Büchi and co-Büchi objectives w.r.t. X.

$$\mathtt{O} \stackrel{\text{def}}{=} \mathsf{FG}X \ \vee \ ((\mathsf{GF}X) \wedge \mathtt{O}_1)$$

The objective $\mathtt{O}$ is shift-invariant, but neither submixing nor inverse-submixing.

▶ **Proposition 5.21.** *Max has optimal MR (and FDD) strategies for* $\mathtt{O}$ *in maximizing MDPs.*

Proof. Given a maximizing MDP, we can consider the end components wrt. an optimal strategy. For each winning end component E there are several possible cases.

If $\mathsf{FG}X$ is satisfied almost surely in E then an MD strategy is sufficient inside E, since that is a co-Büchi objective.

Otherwise, if it is possible to satisfy $((\mathsf{GF}X) \wedge \mathtt{O}_1)$ almost surely in E, then there are two cases. In the first case, it is possible to satisfy even the stronger objective $((\mathsf{GF}X) \wedge \mathtt{MP} > 0)$ almost surely in E. This objective is a special case of $\mathtt{MP} > 0 \cap \mathtt{EPAR}$ and thus an MR (or alternatively FDD) strategy is sufficient by Theorem 5.3. In the second case, it is possible to almost surely satisfy $((\mathsf{GF}X) \wedge \mathtt{O}_1)$ but not $((\mathsf{GF}X) \wedge \mathtt{MP} > 0)$ inside E. There must exist a state $x \in X$ inside E, such that one can satisfy $((\mathsf{GF}x) \wedge \mathtt{O}_1)$. Since E is finite, it must be possible to satisfy $((\mathsf{GF}x) \wedge \mathtt{MP} = 0)$ almost surely. Thus, there must exist a strategy from state x such that x is re-visited almost surely with an expected total reward $= 0$. Playing such a strategy repeatedly is also sufficient to satisfy $((\mathsf{GF}x) \wedge \mathtt{O}_1)$ almost

surely. Thus it suffices to play an MD strategy for the optimal expected total reward (paid out upon visiting x) inside E.

Hence inside every winning end component an MD strategy or an MR (resp. FDD) strategy suffices. Elsewhere, we can fix an MD strategy that maximizes the chance of reaching a winning end component. Altogether this yields an optimal MR strategy (resp. FDD strategy). ◀

▶ **Proposition 5.22.** *Min has optimal MR (and FDD) strategies for* $\mathtt{O}$ *in minimizing MDPs.*

Proof. It suffices to show the existence of optimal Max MR (and FDD) strategies for the complement objective $\overline{\mathtt{O}} = \mathsf{GF}\overline{X} \ \wedge \ ((\mathsf{FG}\overline{X}) \vee \overline{\mathtt{O}_1}) = (\mathsf{FG}\overline{X}) \vee (\mathsf{GF}\overline{X} \wedge \overline{\mathtt{O}_1})$.

Like in the proof of Proposition 5.21, we can consider the strategies in winning end components E. For the co-Büchi objective $\mathsf{FG}\overline{X}$, an MD strategy in E suffices. Otherwise, in finite-state MDPs, we can win $(\mathsf{GF}\overline{X} \wedge \overline{\mathtt{O}_1})$ almost surely iff we can win $(\mathsf{GF}\overline{X} \wedge \mathtt{MP} < 0)$ almost surely. For this an MR (or alternatively FDD) strategy is sufficient by Theorem 5.3.

Hence inside every winning end component an MD strategy or an MR (resp. FDD) strategy suffices. Elsewhere, we can fix an MD strategy that maximizes the chance of reaching a winning end component. Altogether this yields an optimal MR strategy (resp. FDD strategy). ◀

However, in deterministic 2-player games, optimal Max strategies for $\mathtt{O}$ require infinite memory.

▶ **Proposition 5.23.** *Consider the deterministic 2-player game* $\mathcal{G}$ *from Figure 5.4, and objective* $\mathtt{O}$ *with* $X \stackrel{def}{=} \{s_0\}$*. Max can win objective* $\mathtt{O}$ *surely, but every FR Max strategy cannot guarantee any positive probability for* $\mathtt{O}$*.*

Proof. Towards the first part, we define an optimal Max strategy that keeps track of the total reward and plays as follows. Whenever s_1 is reached with some negative total reward $-n$, Max loops $n-1$ times at s_1 and then goes back to s_0, which makes the total reward 0. This ensures objective $\mathtt{O}$ against any Min strategy as follows. In plays where Min eventually stays in s_0 forever we have $\mathsf{FG}X$ and thus $\mathtt{O}$. Otherwise, s_0 and s_1 are both visited infinitely often and the lim sup of the total reward is 0, and thus $\mathtt{O}$ holds.

Towards the second part, we consider an FR Max strategy σ with m memory modes (allowing randomized updates), starting in memory mode m_0. We will construct a Min strategy π such that $\mathcal{P}^{\mathcal{G}}_{\sigma,\pi,s_0}(\mathtt{O}) = 0$.

For every memory mode m let $p(\mathsf{m})$ be probability that, in state s_1 and memory mode m, in the next step σ returns to s_0. Let $p \stackrel{\text{def}}{=} \min\{p(\mathsf{m}) \mid p(\mathsf{m}) > 0\} > 0$. Let π be the MR Min strategy that in s_0 goes to s_1 with probability $p/2$ and to s_0 with probability $1 - p/2$. Fixing the strategies σ, π yields a Markov chain with $2m$ states and initial state (s_0, m_0).

Let E be the event that s_1 is visited when σ is in some memory mode m with $p(\mathsf{m}) = 0$.

Conditioned under E, the properties $\mathsf{FG}X$ and $\mathsf{GF}X$ are not satisfied and thus $\mathtt{O}$ is not satisfied, *i.e.*, $\mathcal{P}^{\mathcal{G}}_{\sigma,\pi,s_0}(\mathtt{O} \cap E) = 0$. If $\overline{E}$ is a null set then this shows the required property.

Otherwise, conditioned under $\overline{E}$, both states s_0 and s_1 are visited infinitely often almost surely, since $p > p/2 > 0$. In the steady state, the probability α of being in s_0 satisfies $\alpha \geq \alpha(1 - p/2) + (1 - \alpha)p$, and therefore $\alpha \geq 2/3$. Hence the (conditional under $\overline{E}$) expected mean payoff is $\leq \alpha(-1) + (1 - \alpha) \leq -1/3$. It follows that $\mathcal{P}^{\mathcal{G}}_{\sigma,\pi,s_0}(\mathtt{MP} < 0 \mid \overline{E}) = 1$ and thus $\mathcal{P}^{\mathcal{G}}_{\sigma,\pi,s_0}(\mathtt{O}_1 \mid \overline{E}) = 0$. Moreover we have $\mathcal{P}^{\mathcal{G}}_{\sigma,\pi,s_0}(\mathsf{FG}X) = 0$, since $p/2 > 0$ and thus $\mathcal{P}^{\mathcal{G}}_{\sigma,\pi,s_0}(\mathtt{O} \cap \overline{E}) = 0$.

Finally, $\mathcal{P}^{\mathcal{G}}_{\sigma,\pi,s_0}(\mathtt{O}) = \mathcal{P}^{\mathcal{G}}_{\sigma,\pi,s_0}(\mathtt{O} \cap E) + \mathcal{P}^{\mathcal{G}}_{\sigma,\pi,s_0}(\mathtt{O} \cap \overline{E}) = 0$. ◀

Chapter 6

Summary & Outlook

In this chapter, we briefly summarize the results presented in this thesis and consider possible future questions or research directions that can improve or extend the present results.

6.1 Energy-Parity

Summary

We gave a procedure to compute ε-approximations of the value of combined energy-parity objectives in SSGs. The decidability of questions about the exact values is open, but the problem is at least as hard as the positivity problem for linear recurrence sequences [Pir21, Section 5.2.3].

Unlike almost surely winning Max strategies which require infinite memory in general [MSTW17, MSTW21], ε-optimal strategies for either player require only finite memory with at most doubly exponentially many memory modes for unary rewards.

Further questions and possibilities

An interesting topic for further study is whether these results can be extended to other combined objectives where the parity part is replaced by something else, *i.e.*, energy-X for some objective X (e.g., some other color-based condition like Rabin/Streett, or a quantitative objective about multidimensional transition rewards). While our proofs are not completely specific to parity, they do use many strong properties that parity satisfies.

- Shift-invariance of $\texttt{EPAR}$ is used in several places, e.g. in Lemma 3.5 (and thus its consequences) and for the correctness of the constructions in Section 3.6.
- We use the fact that $\texttt{EPAR}$ goes well together with $\texttt{LimInf}(> -\infty)$, *i.e.*, the objective $\texttt{Gain} = \texttt{LimInf}(> -\infty) \cap \texttt{EPAR}$ allows optimal FDD strategies for Max in MDPs; cf. Lemma 3.2.
- The submixing property of $\texttt{OPAR} = \overline{\texttt{EPAR}}$ is used in Theorem 3.4 to lift Lemma 3.2 from MDPs to SSGs.

A second direction to explore is to decrease the complexity of the given procedure in Theorem 3.1. The computational complexity of 3-NEXPTIME for binary rewards can be seen as coming from 4 main sources.

1. The first exponential is a bound for the N in rising MDPs with unary rewards.
2. When the rewards are binary, the size of the intermediate rising MDP $\mathcal{M}'$ constructed will increase by an exponential.
3. A third exponential comes from the fact that optimal *deterministic* strategies for $\texttt{Gain} = \texttt{LimInf}(-\infty) \cap \texttt{EPAR}$ provably might need exponential memory and as we consider the minimizing MDP induced by fixing the optimal strategy for $\texttt{Gain}$, this results in a third exponential blowup.
4. Finally, we need non-determinism because there is currently no known polynomial time algorithm to solve parity games.

Solving Item 4. would certainly help with resolving non-determinism, but this is a long-standing open problem even in the case of deterministic 2-player games. Item 1. seems tight, *i.e.*, there could be a family of One-counter MDPs or even Markov chains where N must be exponential in the size of the MDP to get to an ε-approximation. Items 2. and 3. seem promising to tackle because we are currently using the result for unary case as a black box for binary rewards as well which could potentially be improved. And if we consider *randomized* optimal strategies for $\texttt{Gain}$, this could potentially be of polynomial size as well.

6.2 Energy-MeanPayoff

Summary

We showed that deterministic finite memory suffices to win almost surely the Energy-MeanPayoff objective, which requires the energy condition to be satisfied on the first dimension and achieve a strictly positive meanpayoff on the remaining dimensions. We also show that exponential memory is both necessary and sufficient, where the lower bound holds even for randomized strategies and in games where the rewards are from $\{-1, 0, 1\}$.

Further questions and possibilities

Similar to Chapter 3, one could consider the problem of approximating the value in MDPs or stochastic 2-player games. The hardness results for computing the exact value again follows from [Pir21]. A direct extension would be to look at the strategy complexity for the Energy-MeanPayoff objective in games and whether the finite memory result still holds.

6.3 Lifting and MeanPayoff-Parity

Summary

While finite-memory Max strategies for shift-invariant inverse-submixing objectives can be lifted from MDPs to 2-player stochastic games, this requires exponentially many extra memory modes. This extra memory cannot be avoided in general, even when using randomized strategies (Theorem 5.2).

The mean-payoff-parity objective $\texttt{MP} > 0 \cap \texttt{EPAR}$ in Section 5.4 provides an interesting example where randomization in strategies drastically reduces the amount of memory required (from exponential to polynomial), but does not eliminate the need for memory entirely.

Finally, our counterexamples in Section 5.5 show that the different method of [GZ09, Theorem 9] to lift deterministic strategies from MDPs to 2-player stochastic games cannot be generalized to randomized strategies.

Further questions and possibilities

Regarding the class of shift-invariant inverse-submixing objectives a possible open question is concerned with the exact trade-off between memory and attainment, *i.e.*, how good can randomized strategies with a suboptimal number of memory modes be in the worst case, relative to strategies with sufficient memory.

Another direction concerning $\texttt{MP} > 0 \cap \texttt{EPAR}$ is to further understand the type of strategies required for almost sure satisfaction. Are randomized updates strictly necessary for strategies with memory modes at most number of even colors? What if we relax the memory requirements and allow polynomial memory? Can we get away with using randomization only in choosing the successor state?

A third possibility is to have a characterization for the minimum number of memory modes required to win $\texttt{MP} > 0 \cap \texttt{EPAR}$ almost surely in a game where it is known that Max can almost surely win $\texttt{MP} > 0 \cap \texttt{EPAR}$. The bound $\#(\text{distinct even colors})$, while an optimal complexity over all instances, is not optimal for every instance. It is unclear if the strategy derived in Section 5.4 is the one with the least possible amount of memory required for every game or if there is a different algorithm that can do at least as good and strictly better on certain instances. As the strategy we suggested is based on attractor decompositions, one could parametrize the game in terms of *Strahler number* of the game instead of number of even colors and ask: Is the minimum number of memory modes required to almost surely win $\texttt{MP} > 0 \cap \texttt{EPAR}$ in a game with Strahler number k, exactly k? The Strahler number as defined in [DJT20] is only defined for deterministic 2-player games. One can nevertheless formulate a definition in stochastic 2-player games based on *positive* attractor decompositions or simply use known reductions [CJH03, Figure 1] which take an almost surely winning region of Max for $\texttt{EPAR}$ and convert it into a deterministic 2-player game where every vertex is winning for Max and define the Strahler number for the original game in terms of the Strahler number for the constructed one.

Bibliography

[ABE+18] Mohammed Alshiekh, Roderick Bloem, Rüdiger Ehlers, Bettina Könighofer, Scott Niekum, and Ufuk Topcu. Safe reinforcement learning via shielding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[BBC+14] Tomás Brázdil, Václav Brožek, Krishnendu Chatterjee, Vojtěch Forejt, and Antonín Kučera. Markov decision processes with multiple long-run average objectives. *Logical Methods in Computer Science*, 10, 2014.

[BBE10a] Tomás Brázdil, Václav Brozek, and Kousha Etessami. One-Counter Stochastic Games. In Kamal Lodaya and Meena Mahajan, editors, *IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science (FSTTCS 2010)*, volume 8 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 108–119, Dagstuhl, Germany, 2010. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik. Full version at `http://arxiv.org/abs/1009.5636`.

[BBE10b] Tomás Brázdil, Václav Brozek, and Kousha Etessami. One-counter stochastic games. *CoRR*, abs/1009.5636, 2010.

[BBE+10c] Tomás Brázdil, Václav Brožek, Kousha Etessami, Antonín Kučera, and Dominik Wojtczak. One-counter Markov decision processes. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 863–874, 2010.

[BBEK13] T. Brázdil, V. Brožek, K. Etessami, and A. Kučera. Approximating the Termination Value of One-Counter MDPs and Stochastic Games. *Information and Computation*, 222:121–138, 2013.

[BCJ18] Roderick Bloem, Krishnendu Chatterjee, and Barbara Jobstmann. Graph Games and Reactive Synthesis. In *Handbook of Model Checking*, pages 921–962. Springer International Publishing, 2018.

[BCKT18] Tomáš Brázdil, Krishnendu Chatterjee, Jan Křetínskỳ, and Viktor Toman. Strategy representation by decision trees in reactive synthesis. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 385–407. Springer, 2018.

[BFRR17] Véronique Bruyère, Emmanuel Filiot, Mickael Randour, and Jean-François Raskin. Meet your expectations with guarantees: Beyond worst-case synthesis in quantitative games. *Information and Computation*, 254:259–295, 2017. SR 2014.

[BHRR19] Véronique Bruyère, Quentin Hautem, Mickael Randour, and Jean-François Raskin. Energy Mean-Payoff Games. In Wan Fokkink and Rob van Glabbeek, editors, *30th International Conference on Concurrency Theory (CONCUR 2019)*, volume 140 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 21:1–21:17, Dagstuhl, Germany, 2019. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[Bil08] Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

[BKK14] Tomás Brázdil, Stefan Kiefer, and Antonín Kučera. Efficient analysis of probabilistic programs with an unbounded counter. *Journal of the ACM*, 61(6):41:1–41:35, 2014.

[BKN16] T. Brázdil, A. Kučera, and P. Novotný. Optimizing the expected mean payoff in energy Markov decision processes. In *International Symposium on Automated Technology for Verification and Analysis (ATVA)*, volume 9938 of *LNCS*, pages 32–49, 2016.

[BORV23] Patricia Bouyer, Youssouf Oualhadj, Mickael Randour, and Pierre Vandenhove. Arena-independent finite-memory determinacy in stochastic games. *Logical Methods in Computer Science*, 19, 2023.

[CD10] Krishnendu Chatterjee and Laurent Doyen. Energy parity games. In *International Colloquium on Automata, Languages and Programming (ICALP)*, volume 6199 of *LNCS*, pages 599–610, 2010.

[CD11a] Krishnendu Chatterjee and Laurent Doyen. Energy and mean-payoff parity Markov decision processes. In *International Symposium on Mathematical Foundations of Computer Science (MFCS)*, volume 6907, pages 206–218, 2011.

[CD11b] Krishnendu Chatterjee and Laurent Doyen. Games and Markov decision processes with mean-payoff parity and energy parity objectives. In *Mathematical and Engineering Methods in Computer Science (MEMICS)*, volume 7119 of *LNCS*, pages 37–46. Springer, 2011.

[CDAHS03] Arindam Chakrabarti, Luca De Alfaro, Thomas A Henzinger, and Mariëlle Stoelinga. Resource interfaces. In *International Workshop on Embedded Software*, pages 117–133, 2003.

[CDGH15] Krishnendu Chatterjee, Laurent Doyen, Hugo Gimbert, and Thomas A. Henzinger. Randomness for free. *Information and Computation*, 245:3 – 16, 2015.

[CDGO14] Krishnendu Chatterjee, Laurent Doyen, Hugo Gimbert, and Youssouf Oualhadj. Perfect-information stochastic mean-payoff parity games. In *International Conference on Foundations of Software Science and Computational Structures (FoSSaCS)*, volume 8412 of *LNCS*, 2014.

[CGP99] E.M. Clarke, O. Grumberg, and D. Peled. *Model Checking*. MIT Press, Dec. 1999.

[Cha07] Krishnendu Chatterjee. Optimal strategy synthesis in stochastic Müller games. In *International Conference on Foundations of Software Science and Computational Structures (FoSSaCS)*, volume 4423 of *Lecture Notes in Computer Science*, pages 138–152. Springer, 2007.

[CHJ05] Krishnendu Chatterjee, Thomas A Henzinger, and Marcin Jurdzinski. Mean-payoff parity games. In *Logic in Computer Science (LICS)*, pages 178–187, 2005.

[CHP07] Krishnendu Chatterjee, Thomas A. Henzinger, and Nir Piterman. Generalized parity games. In Helmut Seidl, editor, *International Conference on Foundations of Software Science and Computational Structures (FoSSaCS)*, volume 4423 of *LNCS*, pages 153–167. Springer, 2007.

[CJH03] Krishnendu Chatterjee, Marcin Jurdziński, and Thomas A. Henzinger. Simple stochastic parity games. In *Computer Science Logic (CSL)*, volume 2803 of *LNCS*, pages 100–113. Springer, 2003.

[CKK17] Krishnendu Chatterjee, Zuzana Kretínská, and Jan Kretínský. Unifying two views on multiple mean-payoff objectives in Markov decision processes. *Logical Methods in Computer Science*, 13(2), 2017.

[CMH06] Krishnendu Chatterjee, Rupak Majumdar, and Thomas A Henzinger. Markov decision processes with multiple objectives. In *Annual symposium on theoretical aspects of computer science*, pages 325–336. Springer, 2006.

[Con92] Anne Condon. The complexity of stochastic games. *Information and Computation*, 96(2):203–224, 1992.

[CR15] Lorenzo Clemente and Jean-Francois Raskin. Multidimensional beyond worst-case and almost-sure problems for mean-payoff objectives. In *Proceedings of the 2015 30th Annual ACM/IEEE Symposium on Logic in Computer Science (LICS)*, page 257–268, 2015.

[CRR14] Krishnendu Chatterjee, Mickael Randour, and Jean-François Raskin. Strategy synthesis for multi-dimensional quantitative objectives. *Acta Informatica*, 51(3-4):129–163, 2014.

[DA97] Luca De Alfaro. *Formal verification of probabilistic systems*. PhD thesis, Stanford University, 1997.

[DJL18] Laure Daviaud, Martin Jurdziński, and Ranko Lazić. A pseudo-quasi-polynomial algorithm for mean-payoff parity games. In *Logic in Computer Science (LICS)*, pages 325–334, 2018.

[DJT20] Laure Daviaud, Marcin Jurdziński, and K. S. Thejaswini. The Strahler Number of a Parity Game. In *47th International Colloquium on Automata, Languages, and Programming (ICALP 2020)*, volume 168 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 123:1–123:19, Dagstuhl, Germany, 2020. Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

[DM23] Mohan Dantam and Richard Mayr. Approximating the value of energy-parity objectives in simple stochastic games. In *MFCS*, volume 272 of *LIPIcs*, pages 38:1–38:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2023.

[DM24] Mohan Dantam and Richard Mayr. Finite-memory strategies for almost-sure energy-meanpayoff objectives in mdps. In *ICALP*, volume 297 of *LIPIcs*, pages 133:1–133:17. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2024.

[EKVY08] Kousha Etessami, Marta Kwiatkowska, Moshe Y. Vardi, and Mihalis Yannakakis. Multi-objective model checking of Markov decision processes. *Logical Methods in Computer Science*, 4, 2008.

[EOS07] Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. Internet advertising and the generalized second-price auction. *American Economic Review*, 97(1):242–259, 2007.

[EvdPSW03] Graham Everest, Alfred Jacobus van der Poorten, Igor Shparlinski, and Thomas Ward. *Recurrence sequences*. ACM, 2003.

[Gel14] Marcus Gelderie. *Strategy machines: representation and complexity of strategies in infinite games*. PhD thesis, Aachen, Techn. Hochsch., Diss., 2014, 2014.

[GH10] Hugo Gimbert and Florian Horn. Solving simple stochastic tail games. In *ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 847–862, 2010.

[Gil57] Dean Gillette. Stochastic games with zero stop probabilities. *Contributions to the Theory of Games*, 3:179–187, 1957.

[GK23] H. Gimbert and E. Kelmendi. Submixing and shift-invariant stochastic games. *International Journal of Game Theory*, 52:1179–1214, 2023.

[Goe94] Michael X. Goemans. An introduction to linear programming. `https://www.cs.cmu.edu/afs/cs/user/glmiller/public/Scientific-Computing/F-11/RelatedWork/Goemans-LP-notes.pdf`, 1994.

[GOP11] Hugo Gimbert, Youssouf Oualhadj, and Soumya Paul. Computing optimal strategies for Markov decision processes with parity and positive-average conditions. working paper or preprint, 2011.

[GZ09] Hugo Gimbert and Wieslaw Zielonka. Pure and Stationary Optimal Strategies in Perfect-Information Stochastic Games. working paper or preprint, December 2009.

[HJK+22] Christian Hensel, Sebastian Junges, Joost-Pieter Katoen, Tim Quatmann, and Matthias Volk. The probabilistic model checker STORM. *International Journal on Software Tools for Technology Transfer (STTT)*, 24(4):589–610, 2022.

[Hor09] Florian Horn. Random fruits on the Zielonka tree. In *International Symposium on Theoretical Aspects of Computer Science (STACS)*, volume 3 of *LIPIcs*, pages 541–552. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, Germany, 2009.

[JLS15] Marcin Jurdziński, Ranko Lazić, and Sylvain Schmitz. Fixed-dimensional energy games are in pseudo-polynomial time. In *International Colloquium on Automata, Languages and Programming (ICALP)*, volume 9135, pages 260–272, 2015.

[Jur98a] M. Jurdziński. Deciding the winner in parity games is in UP ∩ co-UP. *Information Processing Letters*, 68:119–124, 1998.

[Jur98b] Marcin Jurdziński. Deciding the winner in parity games is in UP ∩ co-UP. *Information Processing Letters*, 68(3):119–124, 1998.

[KNP11] Marta Kwiatkowska, Gethin Norman, and David Parker. PRISM 4.0: Verification of probabilistic real-time systems. In *Proceedings of the 23rd International Conference on Computer Aided Verification (CAV)*, pages 585–591. Springer, 2011.

[Kop08] Eryk Kopczyński. *Half-positional Determinacy of Infinite Games*. PhD thesis, Warsaw University, 2008.

[Mar98] Donald A. Martin. The determinacy of Blackwell games. *Journal of Symbolic Logic*, 63(4):1565–1581, 1998.

[Mea55] George H. Mealy. A method for synthesizing sequential circuits. *The Bell System Technical Journal*, 34(5):1045–1079, 1955.

[Mil00] Paul Milgrom. Putting auction theory to work: The simultaneous ascending auction. *Journal of Political Economy*, 108(2):245–272, 2000.

[MR22] James C. A. Main and Mickael Randour. Different strokes in randomised strategies: Revisiting Kuhn's theorem under finite-memory assumptions. In *CONCUR*, volume 243 of *LIPIcs*, pages 22:1–22:18. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2022.

[MS03] A. Maitra and W. Sudderth. Stochastic games with Borel payoffs. In *Stochastic Games and Applications*, pages 367–373. Kluwer, Dordrecht, 2003.

[MSTW17] Richard Mayr, Sven Schewe, Patrick Totzke, and Dominik Wojtczak. MDPs with Energy-Parity Objectives. In *Logic in Computer Science (LICS)*. IEEE, 2017.

[MSTW21] Richard Mayr, Sven Schewe, Patrick Totzke, and Dominik Wojtczak. Simple stochastic games with almost-sure energy-parity objectives are in NP and coNP. In *Proc. of Fossacs*, volume 12650 of *LNCS*, 2021. Extended version on arXiv.

[NRTV07] Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V Vazirani. *Algorithmic Game Theory*. Cambridge University Press, 2007.

[OW15] Joël Ouaknine and James Worrell. On linear recurrence sequences and loop termination. *ACM SIGLOG News*, 2(2):4–13, 2015.

[PB20] Jakob Piribauer and Christel Baier. On Skolem-Hardness and Saturation Points in Markov Decision Processes. In Artur Czumaj, Anuj Dawar, and Emanuela Merelli, editors, *Proc. of ICALP*, volume 168 of *LIPIcs*, pages 138:1–138:17, Dagstuhl, Germany, 2020. Schloss Dagstuhl–Leibniz-Zentrum für Informatik.

[PB23] Jakob Piribauer and Christel Baier. Positivity-hardness results on Markov decision processes. working paper or preprint, 2023.

[Pir21] Jakob Piribauer. *On non-classical stochastic shortest path problems*. PhD thesis, Technische Universität Dresden, Germany, 2021.

[PR89] A. Pnueli and R. Rosner. On the synthesis of a reactive module. In *Annual Symposium on Principles of Programming Languages (POPL)*, pages 179–190, 1989.

[Pur95] A. Puri. *Theory of hybrid systems and discrete event structures*. PhD thesis, University of California, Berkeley, 1995.

[Put94] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994.

[Rot84] Alvin E Roth. The evolution of the labor market for medical interns and residents: a case study in game theory. *Journal of Political Economy*, 92(6):991–1016, 1984.

[RSÜ04] Alvin E Roth, Tayfun Sönmez, and M Utku Ünver. Kidney exchange. *The Quarterly Journal of Economics*, 119(2):457–488, 2004.

[RW87] Peter J. Ramadge and W. Murray Wonham. Supervisory control of a class of discrete event processes. *SIAM journal on control and optimization*, 25(1):206–230, 1987.

[SB18] R.S. Sutton and A.G Barto. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning. MIT Press, 2018.

[Sch02] Manfred Schäl. Markov decision processes in finance and dynamic options. In *Handbook of Markov Decision Processes*, pages 461–487. Springer, 2002.

[Sha53] Lloyd S. Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

[Van23] Pierre Vandenhove. *Strategy complexity of zero-sum games on graphs*. PhD thesis, Université Paris-Saclay; Université de Mons, 2023.

[Var07] Hal R Varian. Position auctions. *International Journal of Industrial Organization*, 25(6):1163–1178, 2007.

[Zie98] W. Zielonka. Infinite games on finitely coloured graphs with applications to automata on infinite trees. *Theoretical Computer Science*, 200(1-2):135–183, 1998.