Strategy Complexity of Parity Objectives in Countable MDPs

Strategy Complexity of Parity Objectives in Countable MDPs

Stefan Kiefer
Department of Computer Science, University of Oxford, UK

Richard Mayr
School of Informatics, University of Edinburgh, UK

Mahsa Shirmohammadi
CNRS & IRIF, Université de Paris, FR

Patrick Totzke
Department of Computer Science, University of Liverpool, UK

Abstract

We study countably infinite MDPs with parity objectives. Unlike in finite MDPs, optimal strategies need not exist, and may require infinite memory if they do. We provide a complete picture of the exact strategy complexity of ε-optimal strategies (and optimal strategies, where they exist) for all subclasses of parity objectives in the Mostowski hierarchy. Either MD-strategies, Markov strategies, or 1-bit Markov strategies are necessary and sufficient, depending on the number of colors, the branching degree of the MDP, and whether one considers ε-optimal or optimal strategies. In particular, 1-bit Markov strategies are necessary and sufficient for ε-optimal (resp. optimal) strategies for general parity objectives.

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

Keywords and phrases Markov decision processes, Parity objectives, Levy's zero-one law

Digital Object Identifier 10.4230/LIPIcs.CONCUR.2020.39

Related Version A full version of the paper is [13], available at `http://arxiv.org/abs/2007.05065`.

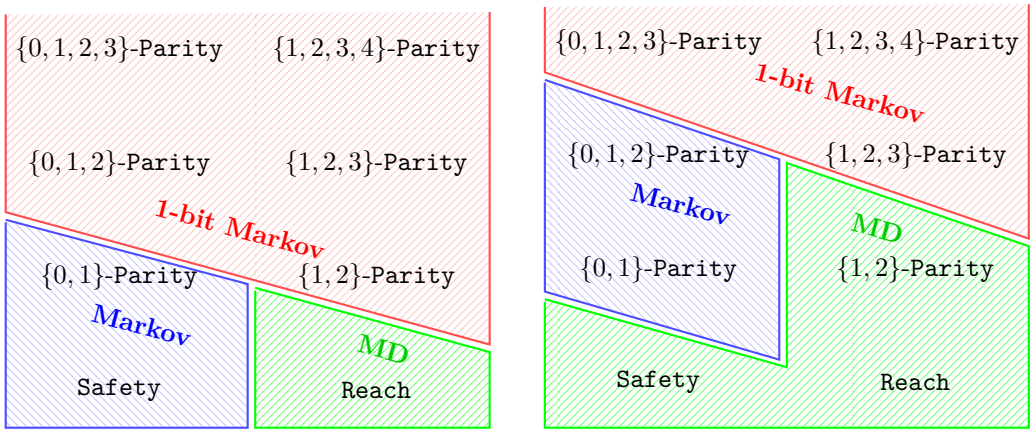
Funding *Stefan Kiefer*: Supported by a Royal Society University Fellowship.

1 Introduction

Background. Markov decision processes (MDPs) are a standard model for dynamic systems that exhibit both stochastic and controlled behavior [17]. MDPs play a prominent role in numerous domains, including artificial intelligence and machine learning [20, 19], control theory [4, 1], operations research and finance [5, 18], and formal verification [7, 2].

An MDP is a directed graph where states are either random or controlled. Its observed behavior is described by runs, which are infinite paths that are, in part, determined by the choices of a controller. If the current state is random then the next state is chosen according to a fixed probability distribution. Otherwise, if the current state is controlled, the controller can choose a distribution over all possible successor states. By fixing a strategy for the controller (and initial state), one obtains a probability space of runs of the MDP. The goal of the controller is to optimize the expected value of some objective function on the runs.

The type of strategy necessary to achieve an optimal (resp. ε-optimal) value for a given objective is called its *strategy complexity*. There are different types of strategies, depending on whether one can take the whole history of the run into account (history-dependent; (H)), or whether one is limited to a finite amount of memory (finite memory; (F)) or whether decisions are based only on the current state (memoryless; (M)). Moreover, the strategy type depends on whether the controller can randomize (R) or is limited to deterministic



(a) ε-optimal strategies for infinitely branching MDPs.

(b) Optimal strategies for infinitely branching MDPs.

Figure 1 These diagrams show the strategy complexity of ε-optimal strategies and optimal strategies (where they exist) for parity objectives. Depending on the position in the Mostowski hierarchy, either MD-strategies (green), deterministic Markov-strategies (blue) or deterministic 1-bit Markov strategies (red) are necessary and sufficient (and randomization does not help [12]). If the MDPs are finitely branching then the Markov strategies can be replaced by MD-strategies (i.e., the blue parts turn green), but the deterministic 1-bit Markov part (red) remains unchanged.

choices (D). The simplest type, MD, refers to memoryless deterministic strategies. *Markov strategies* are strategies that base their decisions only on the current state and the number of steps in the history of the run. Thus they do use infinite memory, but only in a very restricted form by maintaining an unbounded step-counter. Slightly more general are *1-bit Markov strategies* that use 1 bit of extra memory in addition to a step-counter.

Parity objectives. We study countably infinite MDPs with parity objectives. Parity conditions are widely used in temporal logic and formal verification, e.g., they can express ω-regular languages and modal μ-calculus [9]. Every state has a *color*, out of a finite set of colors encoded as natural numbers. A run is winning iff the highest color that is seen infinitely often is even. The controller wants to maximize the probability of winning runs. The Mostowski hierarchy [15] is a classification of parity conditions based on restricting the set of allowed colors. For instance, $\{1, 2, 3\}$-`Parity` objectives only use colors 1, 2, and 3. This includes Büchi ($\{1, 2\}$-`Parity`) and co-Büchi objectives ($\{0, 1\}$-`Parity`), both of which further subsume reachability and safety objectives.

Related work. In *finite* MDPs, there always exist optimal MD-strategies for parity objectives. In fact, this holds even for finite turn-based 2-player stochastic parity games [6, 23]. Similarly, there always exist optimal MD-strategies in countably infinite *non-stochastic* turn-based 2-player parity games [22].

The picture is more complex for countably infinite MDPs. Optimal strategies need not exist (not even for reachability objectives [17, 16]), and ε-optimal strategies for Büchi objectives [10] and optimal strategies for parity objectives [14] require infinite memory.

The paper [14] gave a complete classification whether MD-strategies suffice or whether infinite memory is required for ε-optimal (resp. optimal) strategies for all subclasses of parity objectives in the Mostowski-hierarchy.

However, the mere fact that infinite memory is required for (a subclass of) parity does not establish the precise strategy complexity. E.g., are Markov strategies (or Markov strategies with finite extra memory) sufficient?

In [12] we showed that deterministic 1-bit Markov strategies are both necessary and sufficient for ε-optimal strategies for Büchi objectives. I.e., deterministic 1-bit Markov strategies are sufficient, but neither randomized Markov strategies nor randomized finite-memory strategies are sufficient. This solved a 40-year old problem in gambling theory from [10, 11]. The same paper [12] showed that even for finitely branching MDPs with $\{1,2,3\}$-`Parity` objectives, optimal strategies (where they exist) need to be *at least* deterministic 1-bit Markov in general, i.e., neither randomized Markov nor randomized finite-memory strategies are sufficient.

While the lower bounds for ε-optimal strategies for Büchi objectives (resp. for optimal strategies for $\{1,2,3\}$-`Parity` objectives) carry over to general parity objectives, the upper bounds on the strategy complexity of ε-optimal (resp. optimal) parity remained open.

A basic upper bound and related conjecture. A basic upper bound on the complexity of ε-optimal strategies for parity can be obtained by using a combination of the results of [12] on Büchi objectives (1-bit Markov) and Lévy's zero-one law as follows. (However, note that the following argument does not work directly for optimal strategies.)

Informally speaking, Lévy's zero-one law implies that, for a tail objective (like parity) and any strategy, the level of attainment from the current state almost surely converges to either zero or one. I.e., the runs that always stay in states where the strategy attains something in $(0,1)$ is a null-set. A consequence for parity is that almost all winning runs must eventually, with ever higher probability, commit to winning by some particular color. Thus, with minimal losses (e.g., $\varepsilon/2$), after a sufficiently long finite prefix (depending on ε), one can switch to a strategy that aims to visit some *particular* color x infinitely often. The latter objective is like a Büchi objective where the states of color x are accepting and states of color $> x$ are considered losing sinks. By [12], an $\varepsilon/2$-optimal strategy for such a Büchi objective can be chosen 1-bit Markov. However, one would also need to remember which color x one is supposed to win by and *stick to that color*. The latter is critical, since strategies that switch focus between winning colors infinitely often (e.g., if they follow some local criteria based on the value of the current state wrt. various colors) can end up losing. Overall, the memory needed for such an ε-optimal strategy for parity is: $\lceil \log_2(c) \rceil$ bits for c even colors to remember which color x one is supposed to win by and Markov plus 1 bit for the Büchi strategy (see above), where the Markov step-counter also determines whether one still plays in the prefix. Thus Markov plus $(1 + \lceil \log_2(c) \rceil)$ bits are sufficient. This argument would suggest that more memory is required for more colors. However, our result shows that this is *not* the case.

Our contributions. We show *tight* upper bounds on the strategy complexity of ε-optimal (resp. optimal) strategies for parity objectives: They can be chosen as deterministic 1-bit Markov, regardless of the number of colors. I.e., we provide matching upper bounds to the lower bounds from [12].

In Section 3 we prove Theorem 1. An iterative plastering construction (i.e., fixing player choices on larger and larger subspaces) builds an ε-optimal 1-bit Markov strategy where the probability of never switching between winning even colors is $\geq 1 - \varepsilon$. Its correctness relies heavily on Lévy's zero-one law. The number of iterations is finite and proportional to the number of even colors. It eliminates the need to remember the winning color x and the $\lceil \log_2(c) \rceil$ part of the memory.

▶ **Theorem 1.** *Consider an MDP $\mathcal{M}$, a parity objective and a finite set S_0 of initial states.*
For every $\varepsilon > 0$ there exists a deterministic 1-bit Markov strategy that is ε-optimal from every state $s \in S_0$.

In Section 4 we prove Theorem 2. If an optimal strategy exists, then an optimal 1-bit Markov strategy can be constructed by the so-called *sea urchin* construction. It is a very complex plastering construction with infinitely many iterations that uses the results of Theorem 1 and Lévy's zero-one law as building blocks. Its name comes from the shape of the subspace in which player choices get fixed: a growing finite body (around a start set S_0) with a finite, but increasing, number of spikes, where each spike is of infinite size; cf. Figure 4. E.g., if the initial states are almost surely winning then, at the stage with i spikes, this strategy attains parity with some probability $\geq 1 - 2^{-i}$ already *inside* this subspace, and in the limit of $i \to \infty$ it attains parity almost surely. A further step even yields a single deterministic 1-bit Markov strategy that is optimal from every state that has an optimal strategy.

▶ **Theorem 2.** *Consider an MDP $\mathcal{M}$ with a parity objective and let S_{opt} be the subset of states that have an optimal strategy.*
There exists a deterministic 1-bit Markov strategy that is optimal from every $s \in S_{opt}$.

In Theorem 1 and Theorem 2 the initial content of the 1-bit memory is irrelevant (cf. Lemma 9, Lemma 18 and Remark 8).

Moreover, we show in Section 5 and Section 6 that in certain subcases deterministic Markov strategies are necessary and sufficient (i.e., these require a Markov step-counter, but not the extra bit): optimal strategies for co-Büchi and $\{0,1,2\}$-`Parity`, and ε-optimal strategies for safety and co-Büchi. In the special case of finitely branching MDPs, these Markov strategies (but not the 1-bit Markov strategies) can be replaced by MD-strategies.

Together with the previously established lower bounds, this yields a complete picture of the *exact* strategy complexity of parity objectives at all levels of the Mostowski hierarchy, for countable MDPs. Figure 1 gives a complete overview.

2 Preliminaries

A *probability distribution* over a countable set S is a function $f : S \to [0,1]$ with $\sum_{s \in S} f(s) = 1$. We write $\mathcal{D}(S)$ for the set of all probability distributions over S.

We study *Markov decision processes* (MDPs) over countably infinite state spaces. Formally, an MDP $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P)$ consists of a countable set S of *states*, which is partitioned into a set $S_\Box$ of *controlled states* and a set $S_\bigcirc$ of *random states*, a *transition relation* $\longrightarrow \subseteq S \times S$, and a *probability function* $P : S_\bigcirc \to \mathcal{D}(S)$. We write $s \longrightarrow s'$ if $(s, s') \in \longrightarrow$, and refer to s' as a *successor* of s. We assume that every state has at least one successor. The probability function P assigns to each random state $s \in S_\bigcirc$ a probability distribution $P(s)$ over its set of successors. A *sink* is a subset $T \subseteq S$ closed under the $\longrightarrow$ relation. An MDP is *acyclic* if the underlying graph $(S, \longrightarrow)$ is acyclic. It is *finitely branching* if every state has finitely many successors and *infinitely branching* otherwise. An MDP without controlled states ($S_\Box = \emptyset$) is a *Markov chain*.

Strategies and Probability Measures. A *run* ρ is an infinite sequence $s_0 s_1 \cdots$ of states such that $s_i \longrightarrow s_{i+1}$ for all $i \in \mathbb{N}$; write $\rho(i) \stackrel{\text{def}}{=} s_i$ for the i-th state along ρ. A *partial run* is a finite prefix of a run. We say that (partial) run ρ *visits* s if $s = \rho(i)$ for some i, and that ρ *starts in* s if $s = \rho(0)$.

A *strategy* is a function $\sigma : S^*S_\Box \to \mathcal{D}(S)$ that assigns to partial runs $\rho s \in S^*S_\Box$ a distribution over the successors of s. A (partial) run $s_0 s_1 \cdots$ is *induced by* strategy σ if for all i either $s_i \in S_\Box$ and $\sigma(s_0 s_1 \cdots s_i)(s_{i+1}) > 0$, or $s_i \in S_\bigcirc$ and $P(s_i)(s_{i+1}) > 0$.

A strategy σ and an initial state $s_0 \in S$ induce a standard probability measure on sets of infinite plays. We write $\mathbb{P}_{\mathcal{M},s_0,\sigma}(\mathcal{R})$ for the probability of a measurable set $\mathcal{R} \subseteq s_0 S^\omega$ of runs starting from s_0. As usual, it is first defined on the *cylinders* $s_0 s_1 \dots s_n S^\omega$, where $s_1, \dots, s_n \in S$: if $s_0 s_1 \dots s_n$ is not a partial run induced by σ then $\mathbb{P}_{\mathcal{M},s_0,\sigma}(s_0 s_1 \dots s_n S^\omega) \stackrel{\text{def}}{=} 0$. Otherwise, $\mathbb{P}_{\mathcal{M},s_0,\sigma}(s_0 s_1 \dots s_n S^\omega) \stackrel{\text{def}}{=} \prod_{i=0}^{n-1} \bar{\sigma}(s_0 s_1 \dots s_i)(s_{i+1})$, where $\bar{\sigma}$ is the map that extends σ by $\bar{\sigma}(ws) = P(s)$ for all $ws \in S^*S_\bigcirc$. By Carathéodory's theorem [3], this extends uniquely to a probability measure $\mathbb{P}_{\mathcal{M},s_0,\sigma}$ on measurable subsets of $s_0 S^\omega$. We will write $\mathbb{E}_{\mathcal{M},s_0,\sigma}$ for the expectation w.r.t. $\mathbb{P}_{\mathcal{M},s_0,\sigma}$. We may drop the subscripts from notations, if it is understood.

Objectives. The objective of the player is determined by a predicate on infinite plays. We assume familiarity with the syntax and semantics of the temporal logic LTL [8]. Formulas are interpreted on the structure $(S, \longrightarrow)$. We use $[\![\varphi]\!]^s \subseteq sS^\omega$ to denote the set of runs starting from s that satisfy the LTL formula φ, which is a measurable set [21]. We also write $[\![\varphi]\!]$ for $\bigcup_{s\in S}[\![\varphi]\!]^s$. Where it does not cause confusion we will identify φ and $[\![\varphi]\!]$ and just write $\mathbb{P}_{\mathcal{M},s,\sigma}(\varphi)$ instead of $\mathbb{P}_{\mathcal{M},s,\sigma}([\![\varphi]\!]^s)$.

Given a set $T \subseteq S$ of states, the *reachability* objective $\texttt{Reach}(T)$ is the set of runs that visit T at least once; and the *safety objective* $\texttt{Safety}(T)$ is the set of runs that never visit T.

Let $\mathcal{C} \subseteq \mathbb{N}$ be a finite set of colors. A *color function* $Col : S \to \mathcal{C}$ assigns to each state s its color $Col(s)$. The parity objective, written as $\texttt{Parity}(Col)$, is the set of infinite runs such that the largest color that occurs infinitely often along the run is even. To define this formally, let $even(\mathcal{C}) = \{i \in \mathcal{C} \mid i \equiv 0 \mod 2\}$. For $\rhd \in \{<, \leq, =, \geq, >\}$, $n \in \mathbb{N}$, and $Q \subseteq S$, let $[Q]^{Col \rhd n} \stackrel{\text{def}}{=} \{s \in Q \mid Col(s) \rhd n\}$ be the set of states in Q with color $\rhd n$. Then

$$\texttt{Parity}(Col) \stackrel{\text{def}}{=} \bigvee_{i \in even(\mathcal{C})} \left(\mathsf{GF}[S]^{Col=i} \wedge \mathsf{FG}[S]^{Col\leq i}\right).$$

The Mostowski hierarchy [15] classifies parity objectives by restricting the range of Col to a set of colors $\mathcal{C} \subseteq \mathbb{N}$. We write $\mathcal{C}$-$\texttt{Parity}$ for such restricted parity objectives. In particular, the classical Büchi and co-Büchi objectives correspond to $\{1,2\}$-$\texttt{Parity}$ and $\{0,1\}$-$\texttt{Parity}$, respectively. These two classes are incomparable but both subsume the reachability and safety objectives. Assuming that T is a sink, $\texttt{Reach}(T) = \texttt{Parity}(Col)$ for the coloring with $Col(s) = 1 \iff s \notin T$ and $\texttt{Safety}(T) = \texttt{Parity}(Col)$ for the coloring with $Col(s) = 1 \iff s \in T$. Similarly, $\{0,1,2\}$-$\texttt{Parity}$ and $\{1,2,3\}$-$\texttt{Parity}$ are incomparable, but they both subsume (modulo renaming of colors) Büchi and co-Büchi objectives.

An objective φ is called a *tail objective* (resp. *suffix-closed*) iff for every run $\rho'\rho$ with some finite prefix ρ' we have $\rho'\rho \in \varphi \Leftrightarrow \rho \in \varphi$ (resp. $\rho'\rho \in \varphi \Rightarrow \rho \in \varphi$). In particular, $\texttt{Parity}(Col)$ is tail for every coloring Col. Moreover, if φ is suffix-closed then $\mathsf{F}\varphi$ is tail.

Strategy Classes. Strategies $\sigma : S^*S_\Box \to \mathcal{D}(S)$ are in general *randomized* (R) in the sense that they take values in $\mathcal{D}(S)$. A strategy σ is *deterministic* (D) if $\sigma(\rho)$ is a Dirac distribution for all partial runs $\rho \in S^*S_\Box$.

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

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

An update function u together with an initial memory m_0 induce a strategy $u[\mathsf{m}_0] : S^*S_\Box \to \mathcal{D}(S)$ as follows. Consider the Markov chain with states set $\mathsf{M} \times S$, transition relation $(\mathsf{M} \times S)^2$ and probability function u. Any partial run $\rho = s_0 \cdots s_i$ in $\mathcal{M}$ gives rise to a set $H(\rho) = \{(\mathsf{m}_0, s_0) \cdots (\mathsf{m}_i, s_i) \mid \mathsf{m}_0, \ldots, \mathsf{m}_i \in \mathsf{M}\}$ of partial runs in this Markov chain. Each $\rho s \in s_0 S^* S_\Box$ induces a probability distribution $\mu_{\rho s} \in \mathcal{D}(\mathsf{M})$, the probability of being in state (m, s) conditioned on having taken some partial run from $H(\rho s)$. We define $u[\mathsf{m}_0]$ such that $u[\mathsf{m}_0](\rho s)(s') \stackrel{\text{def}}{=} \sum_{\mathsf{m},\mathsf{m}' \in \mathsf{M}} \mu_{\rho s}(\mathsf{m}) u((\mathsf{m}, s))(\mathsf{m}', s')$ for all $\rho s \in S^* S_\Box$ and $s' \in S$.

We say that a strategy σ can be *implemented* with memory M (and initial memory m_0) if there exists an update function u such that $\sigma = u[\mathsf{m}_0]$. In this case we may also write $\sigma[\mathsf{m}_0]$ to explicitly specify the initial memory mode m_0. Based on this, we can define several classes of strategies:

- A strategy σ is *memoryless* (M) (also called *positional*) if it can be implemented with a memory of size 1. We may view M-strategies as functions $\sigma : S_\Box \to \mathcal{D}(S)$.
- A strategy σ is *finite memory* (F) if there exists a finite memory M implementing σ. More specifically, a strategy is *k-bit* if it can be implemented with a memory of size 2^k. Such a strategy is then determined by a function $u : \{0,1\}^k \times S \to \mathcal{D}(\{0,1\}^k \times S)$.
- A strategy σ is *Markov* if it can be implemented with the natural numbers $\mathsf{M} = \mathbb{N}$ as the memory, initial memory mode $\mathsf{m}_0 = 0$ and a function u such that the distribution $u(\mathsf{m}, s)$ is over $\{\mathsf{m}+1\} \times S$ for all $\mathsf{m} \in \mathsf{M}$ and $s \in S$. Intuitively, such a strategy depends only on the current state and the number of steps taken so far.
- A strategy σ is *k-bit Markov* if it can be implemented with memory $\mathsf{M} = \mathbb{N} \times \{0,1\}^k$, $\mathsf{m}_0 \in \{0\} \times \{0,1\}^k$ and a function u such that the distribution $u((n, b, s))$ is over $\{n+1\} \times \{0,1\}^k \times S$ for all $(n, b) \in \mathsf{M}$ and $s \in S$.

Deterministic 1-bit strategies are central in this paper; by this we mean strategies that are both deterministic and 1-bit.

Optimal and ε-optimal Strategies. Given an objective φ, the value of state s in an MDP $\mathcal{M}$, denoted by $\mathtt{val}_{\mathcal{M}}(s)$, is the supremum probability of achieving φ. Formally, we have $\mathtt{val}_{\mathcal{M}}(s) \stackrel{\text{def}}{=} \sup_{\sigma \in \Sigma} \mathbb{P}_{\mathcal{M},s,\sigma}(\varphi)$ where Σ is the set of all strategies. For $\varepsilon \geq 0$ and state $s \in S$, we say that a strategy is ε-optimal from s iff $\mathbb{P}_{\mathcal{M},s,\sigma}(\varphi) \geq \mathtt{val}_{\mathcal{M}}(s) - \varepsilon$. A 0-optimal strategy is called optimal. An optimal strategy is almost-surely winning if $\mathtt{val}_{\mathcal{M}}(s) = 1$.

Considering an MD strategy as a function $\sigma : S_\Box \to S$ and $\varepsilon \geq 0$, σ is *uniformly* ε-optimal (resp. uniformly optimal) if it is ε-optimal (resp. optimal) from every $s \in S$.

Fixing and Safe Sets. Let σ be an MD strategy. Given a set $S' \subseteq S$ of states, write $\mathcal{M}[\sigma, S']$ for the MDP obtained from $\mathcal{M}$ by fixing the strategy σ for all states in S', that is, $\mathcal{M}[\sigma, S'] \stackrel{\text{def}}{=} (S, S_\Box \setminus S', S_\bigcirc \cup S', \longrightarrow, P')$ where $P'(s) \stackrel{\text{def}}{=} \sigma(s)$ for all $s \in S'$.

For an objective φ and a threshold $\beta \in [0,1]$, denote by $\mathit{Safe}_{\mathcal{M},\sigma,\varphi}(\beta)$ the set of all states s starting from which σ attains at least probability β; and denote by $\mathit{Safe}_{\mathcal{M},\varphi}(\beta)$ the set of states whose value for φ is at least β. Formally,

$$\mathit{Safe}_{\mathcal{M},\sigma,\varphi}(\beta) \stackrel{\text{def}}{=} \{s \in S \mid \mathbb{P}_{\mathcal{M},s,\sigma}(\varphi) \geq \beta\}, \quad \mathit{Safe}_{\mathcal{M},\varphi}(\beta) \stackrel{\text{def}}{=} \{s \in S \mid \mathtt{val}_{\mathcal{M},\varphi}(s) \geq \beta\}. \quad (1)$$

3 ε-Optimal Strategies for Parity

In this section we prove Theorem 1, stating that ε-optimal strategies for parity objectives can be chosen 1-bit Markov. Given an MDP we convert it by three successive reductions to a structurally simpler MDP where strategies require less sophistication to achieve parity.

First reduction (Finitely Branching). This reduction converts an infinitely branching MDP $\mathcal{M}$ to a finitely branching one $\mathcal{M}'$, with a clear bijection between the strategies in $\mathcal{M}$ and $\mathcal{M}'$. The construction, first presented in our previous work [12], replaces each controlled state s, that has infinitely many successors $(s_i)_{i\in\mathbb{N}}$, with a "ladder" of controlled states $(q_i)_{i\in\mathbb{N}}$, where each q_i has only two successors: q_{i+1} and s_i. Roughly speaking, the controller choice of successor s_n at s in $\mathcal{M}$, is simulated by a series of choices q_{i+1} at q_i, $0 \leq i < n$, followed by a choice of successor s_n in state q_n in $\mathcal{M}'$, and vice versa.

To prevent scenarios when the controller in $\mathcal{M}'$ stays on a ladder and never commits to a decision, we assign color 1 to all states $(q_i)_{i\geq 1}$ on the ladder (q_0 inherits the color of s). Hence, a hesitant run on the ladder is losing for parity. So w.l.o.g. we can assume that the given $\mathcal{M}$ is finitely branching.

▶ **Lemma 3.**

1. *Suppose that for every finitely branching acyclic MDP with a finite set S_0 of initial states, and a parity objective, there exist ε-optimal deterministic 1-bit strategies from S_0.*
   Then even for every infinitely branching acyclic MDP with a finite set S_0 of initial states and a parity objective, there exist ε-optimal deterministic 1-bit strategies from S_0.
2. *Suppose that for every finitely branching acyclic MDP with a parity objective, there exists a deterministic 1-bit strategy that is optimal from all states that have an optimal strategy.*
   Then even for every infinitely branching acyclic MDP with a parity objective, there exists a deterministic 1-bit strategy that is optimal from all states that have an optimal strategy.

Second reduction (Acyclicity). A deterministic 1-bit Markov strategy can be seen as a function $\sigma : \mathbb{N} \times \{0,1\} \times S \to \{0,1\} \times S$, where σ has access to an internal bit $b \in \{0,1\}$, which can be updated freely, and a step counter $k \in \mathbb{N}$, which increments by one in each step. Having b and k, σ produces a decision based on the current state of the MDP.

Following [12], we encode the step-counter from strategies into MDPs s.t. the current state of the system uniquely determines the length of the path taken so far. This translation allows us to focus on acyclic MDPs.

▶ **Lemma 4.** *Consider MDPs with a parity objective and $k \in \mathbb{N}$.*

1. *Suppose that for every acyclic MDP $\mathcal{M}'$ and every* finite *set of initial states S'_0 and $\varepsilon > 0$, there exists a deterministic k-bit strategy that is ε-optimal from all states $s \in S'_0$.*
   Then for every MDP $\mathcal{M}$ and every finite *set of initial states S_0 and $\varepsilon > 0$, there exists a deterministic k-bit Markov strategy that is ε-optimal from all states $s \in S_0$.*
2. *Suppose that for every acyclic MDP $\mathcal{M}'$ and $\varepsilon > 0$, there exists a deterministic k-bit strategy that is ε-optimal from all states. Then for every MDP $\mathcal{M}$ and $\varepsilon > 0$, there exists a deterministic k-bit Markov strategy that is ε-optimal from all states.*
3. *Suppose that for every acyclic MDP $\mathcal{M}'$, where S'_{opt} is the subset of states that have an optimal strategy, there exists a deterministic k-bit strategy that is optimal from all states $s \in S'_{opt}$. Then for every MDP $\mathcal{M}$, where S_{opt} is the subset of states that have an optimal strategy, there exists a deterministic k-bit Markov strategy that is optimal from all states $s \in S_{opt}$.*

By Lemma 4, the sufficiency of deterministic 1-bit strategies in acyclic MDPs implies the sufficiency of deterministic 1-bit Markov strategies in general MDPs. Thus to prove Theorem 1, it suffices to prove the following:

▶ **Theorem 5.** *Consider an acyclic MDP $\mathcal{M}$, a parity objective and a finite set S_0 of states. For every $\varepsilon > 0$ there exists a deterministic 1-bit strategy that is ε-optimal from every $s \in S_0$.*

Third reduction (Layered MDP). This reduction is in the same spirit of the previous one, in which the bit $b \in \{0,1\}$ is transferred from strategies to MDPs. Given an MDP $\mathcal{M}$, the corresponding *layered* MDP $\mathcal{L}(\mathcal{M})$ has two copies of each state $s \in S$ and each transition $t \in \longrightarrow_1$ of $\mathcal{M}$, one augmented with bit 0 and another with bit 1: (s,i) and (t,j) with $i,j \in \{0,1\}$. The states (s,i) are random if $s \in S_\bigcirc$ and controlled if $s \in S_\Box$. All the (t,j) are controlled. If there is a transition $t = (a,b)$ from state a to b in $\mathcal{M}$, there will be two transitions from (a,i) to (t,i), and four transitions from (t,i) to (b,j) in $\mathcal{L}(\mathcal{M})$; see Figure 2.

A 1-bit deterministic strategy in $\mathcal{M}$ at a state a picks a single successor b and may flip the bit from i to j; this is simulated in $\mathcal{L}(\mathcal{M})$ with an MD strategy σ within two consecutive steps: σ first chooses the transition $t = (a,b)$ by $\sigma(a,i) = (t,i)$ and then updates the bit by $\sigma(t,i) = (b,j)$ thereby moving from layer i to layer j. The controlled states (t,i) are essential for a correct simulation, since otherwise the controller cannot freely flip the bit (switch between layers) after it observes the successor chosen randomly at a random state.

▶ **Definition 6** (Layered MDP). *Given an MDP $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow_1, P_1)$ with coloring $Col_1 : S \to \mathcal{C}$, we define the corresponding* layered MDP $\mathcal{L}(\mathcal{M}) = (L, L_\Box, L_\bigcirc, \longrightarrow_2, P_2)$ *with coloring $Col_2 : L \to \mathcal{C}$ as follows.*

- $L \stackrel{def}{=} (S \cup \longrightarrow_1) \times \{0,1\}$ *where the set of controlled states is* $L_\Box \stackrel{def}{=} (S_\Box \cup \longrightarrow_1) \times \{0,1\}$.
- *For all $t \in \longrightarrow_1$ such that $t = (s,s')$ and for all $i,j \in \{0,1\}$, we have:*
  1. $(s,i) \longrightarrow_2 (t,i)$ *and* $(t,i) \longrightarrow_2 (s',j)$,
  2. $P(s,i)((t,i)) \stackrel{def}{=} P(s)(s')$ *iff* $s \in S_\bigcirc$, *and*
  3. $Col_2((s,i)) \stackrel{def}{=} Col_1(s)$ *and* $Col_2((t,i)) \stackrel{def}{=} Col_1(s')$.

The layered MDP of an acyclic MDP is acyclic. For $q \in S \cup \longrightarrow_1$, we refer to the copies of q in layer 0 and layer 1 as *siblings*: $(q,0)$ and $(q,1)$. A set $B \subseteq L$ is *closed* if for each state $(q,i) \in B$ its sibling is also in B. Denote by $Cl(B)$ the minimal closed superset of B.

▶ **Lemma 7.** *Consider an acyclic MDP $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P)$ with a parity objective $\varphi = \mathtt{Parity}(Col)$ and let $\mathcal{L}(\mathcal{M})$ be the corresponding layered MDP.*

For every deterministic 1-bit strategy $u[\mathsf{m}_0]$ in $\mathcal{M}$ there is a corresponding MD strategy τ in $\mathcal{L}(\mathcal{M})$, and vice-versa, such that for every $s_0 \in S$, $\mathbb{P}_{\mathcal{L}(\mathcal{M}),(s_0,\mathsf{m}_0),\tau}(\varphi) = \mathbb{P}_{\mathcal{M},s_0,u[\mathsf{m}_0]}(\varphi)$.

▶ Remark 8. We note that, in a layered system $\mathcal{L}(\mathcal{M})$, any two siblings have the same value w.r.t. a parity objective φ. Moreover, any state s in $\mathcal{M}$ has an optimal strategy iff $(s,0) \in \mathcal{L}(\mathcal{M})$ has an optimal strategy iff its sibling $(s,1)$ has an optimal strategy.

Suppose τ is an MD strategy in $\mathcal{L}(\mathcal{M})$ that is optimal for all states that have an optimal strategy. Let u be the update function of a corresponding 1-bit strategy in $\mathcal{M}$, derived as described in Lemma 7. Then for every state s in $\mathcal{M}$ that has an optimal strategy we have $\mathbb{P}_{\mathcal{M},s,u[0]}(\varphi) = \mathbb{P}_{\mathcal{L}(\mathcal{M}),(s,0),\tau}(\varphi) = \mathbb{P}_{\mathcal{L}(\mathcal{M}),(s,1),\tau}(\varphi) = \mathbb{P}_{\mathcal{M},s,u[1]}(\varphi)$. That is, both $u[0]$ and $u[1]$ are optimal from s, so the initial memory mode is irrelevant. ◀

To prove Theorem 5, given an acyclic MDP, a set of initial states S_0 and $\varepsilon > 0$, we consider the layered MDP $\mathcal{L}(\mathcal{M})$ and set $L_0 = S_0 \times \{0\}$ of initial states. In the following lemma, we prove that there exists a single MD strategy that is ε-optimal starting from every state $\ell_0 \in L_0$ in $\mathcal{L}(\mathcal{M})$. This and Lemma 7 will directly lead to Theorem 5.

▶ **Lemma 9.** *Consider an acyclic MDP $\mathcal{M}$ and parity objective $\varphi = \mathtt{Parity}(Col)$. Let $\mathcal{L}(\mathcal{M})$ be the layered MDP of $\mathcal{M}$ and Col. For all finite sets L_0 of states in $\mathcal{L}(\mathcal{M})$ and all $\varepsilon > 0$ there exists a single MD strategy that is ε-optimal for φ from every state $\ell_0 \in L_0$.*

In the rest of this section, we prove Lemma 9. We fix a layered MDP $\mathcal{L}(\mathcal{M})$ (or simply $\mathcal{L}$) obtained from a given acyclic and finitely branching MDP $\mathcal{M}$ and a coloring $Col : S \to \mathcal{C}$, where the set of states is L and the finite set of initial states is $L_0 \subseteq L$. Let φ be the resulting parity objective in $\mathcal{L}$.

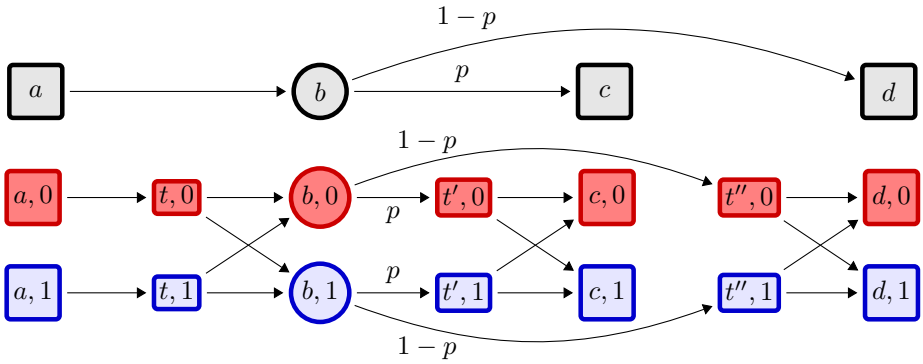


Figure 2 An MDP $\mathcal{M}$ (in grey) and the corresponding layered MDP $\mathcal{L}(\mathcal{M})$ with states of layer 0 and 1 in red and blue, respectively. Here, $t = (a, b)$, $t' = (b, c)$ and $t'' = (b, d)$ are transitions of $\mathcal{M}$.

Recall that $even(\mathcal{C}) = 2\mathbb{N} \cap \mathcal{C}$ denotes the set of even colors. We denote by $e_{\max}$ the largest even color in $even(\mathcal{C})$ and assume w.l.o.g., that $even(\mathcal{C})$ contains all even numbers from 2 to $e_{\max}$ inclusive. We have:

$$\begin{aligned}
\varphi &\stackrel{\text{def}}{=} \bigvee_{e \in even(\mathcal{C})} \left(\mathsf{GF}[L]^{Col=e} \wedge \mathsf{FG}[L]^{Col\leq e}\right) \\
&= \bigvee_{e \in even(\mathcal{C})} \left(\mathsf{FGF}[L]^{Col=e} \wedge \mathsf{FG}[L]^{Col\leq e}\right) \qquad \text{since } \mathsf{GF}[L]^{Col=e} \text{ is a tail objective} \\
&= \bigvee_{e \in even(\mathcal{C})} \mathsf{F}\left(\mathsf{GF}[L]^{Col=e} \wedge \mathsf{G}[L]^{Col\leq e}\right) \qquad \text{since } \mathsf{FG}A \wedge \mathsf{FG}B = \mathsf{F}(\mathsf{G}A \wedge \mathsf{G}B) \\
&= \bigvee_{e \in even(\mathcal{C})} \mathsf{F}\varphi_e \,,
\end{aligned}$$

where $\varphi_e \stackrel{\text{def}}{=} \left(\mathsf{GF}[L]^{Col=e} \wedge \mathsf{G}[L]^{Col\leq e}\right)$. Indeed, φ_e is the set of runs that win through color e (i.e., by visiting color e infinitely often and never visiting larger colors). Since the $\mathsf{F}\varphi_e$ are disjoint, for all states ℓ and strategies σ, we have:

$$\mathbb{P}_{\mathcal{L},\ell,\sigma}(\varphi) = \sum_{e \in even(\mathcal{C})} \mathbb{P}_{\mathcal{L},\ell,\sigma}(\mathsf{F}\varphi_e). \tag{2}$$

Fix $\varepsilon > 0$ and define $\gamma \stackrel{\text{def}}{=} \frac{\varepsilon}{e_{\max}+2}$. To construct an MD strategy $\hat{\sigma}$ that is ε-optimal starting from every state in L_0 we have an iterative procedure. In each iteration, we define $\hat{\sigma}$ at states in some carefully chosen region; and continuing in this fashion, we gradually fix all choices of $\hat{\sigma}$. In an iteration, in order to fix "good" choices in the "right" region we need to carefully observe the behavior of finitely many $\frac{\gamma}{2}$-optimal strategies σ_{ℓ_0}, one for each $\ell_0 \in L_0$, which must respect the choices already fixed in previous iterations. We thus view these strategies σ_{ℓ_0} to be $\frac{\gamma}{2}$-optimal not in $\mathcal{L}$ but in another layered MDP that is derived from $\mathcal{L}$ after fixing the choices of partially defined $\hat{\sigma}$.

In more detail, the proof consists of exactly $\frac{e_{\max}}{2} + 1$ iterations: one iteration for each even color e and a final "reach" iteration. Starting from color 2 and $\mathcal{L}_0 \stackrel{\text{def}}{=} \mathcal{L}$, in the iteration $e \in \{2, \cdots, e_{\max}\}$, we obtain a layered MDP $\mathcal{L}_e$ from $\mathcal{L}_{e-2}$ by fixing a single choice for each controlled state in a set fix_e. Roughly speaking, a run that falls in the set fix_e is likely going to win through φ_e (win through color e). We identify a certain subspace of fix_e, referred to as core_e, such that the following crucial fact holds: Once core_e is visited the run

remains in fix_e with probability at least $1-\gamma$. At the final iteration, we fix the choices of all remaining states to maximize the probability of falling into the union of core_e sets. As mentioned, the majority of such runs that visit core_e, for some color e, will stay in fix_e forever and thus win parity through color e. After all the iterations, all choices of all controlled states are fixed, and this prescribes the MD strategy $\hat{\sigma}$ from L_0 in $\mathcal{L}$.

In order to define the sets fix_e we heavily use Lévy's zero-one law and follow an inductive transformation on objectives. Lévy's zero-one states that, for a given set of (infinite) runs of a Markov chain, if we gradually observe a random run of the chain, we will become more and more certain whether the random run belongs to that set. This law has a strong implications for tail objectives. It asserts that on almost all runs $s_0 s_1 s_2 \cdots$ the limit of the value of s_i w.r.t. a tail objective tends to either 0 or 1.

In each iteration $e \in \{2, \cdots, e_{\max}\}$, we transform an objective ψ_{e-2} to a next objective ψ_e where $\psi_0 \stackrel{\text{def}}{=} \varphi$ is the parity objective and the result of the last transformation is $\psi_{e_{\max}} = \bigvee_{e \in even(\mathcal{C})} \mathsf{F}\mathit{core}_e$. We will also move from the MDP $\mathcal{L}_{e-2}$ to $\mathcal{L}_e$ after the fixings so as to maintain the following **invariant**: For all $\ell_0 \in L_0$, the value of ℓ_0 for ψ_e in $\mathcal{L}_e$ is almost as high as its value for φ in $\mathcal{L}$, that is

$$\mathtt{val}_{\mathcal{L}_e,\psi_e}(\ell_0) \geq \mathtt{val}_{\mathcal{L},\varphi}(\ell_0) - e \cdot \gamma. \tag{3}$$

Recall that $\varphi = \bigvee_{e \in even(\mathcal{C})} \mathsf{F}\varphi_e$. Let $\mathsf{Fix}_0 \stackrel{\text{def}}{=} \emptyset$ and write $\mathsf{Fix}_e \stackrel{\text{def}}{=} \bigcup_{e' \leq e} \mathit{Cl}(\mathit{fix}_{e'})$ for $e \in \{2, 4, \cdots, e_{\max}\}$. We define:

$$\psi_0 \stackrel{\text{def}}{=} \bigvee_{e'>0} \mathsf{F}\varphi_{e'} \wedge \mathsf{G}\,\neg\mathsf{Fix}_0 = \varphi \qquad\qquad \psi_e \stackrel{\text{def}}{=} \bigvee_{e' \leq e} \mathsf{F}\mathit{core}_{e'} \vee \bigvee_{e'>e} (\mathsf{F}\varphi_{e'} \wedge \mathsf{G}\,\neg\mathsf{Fix}_e). \tag{4}$$

At each transformation, we examine the disjunct $\chi_e \stackrel{\text{def}}{=} \mathsf{F}\varphi_e \wedge \mathsf{G}\neg\mathsf{Fix}_{e-2}$ in ψ_{e-2}. The set of runs satisfying this objective χ_e not only win through color e but also avoid the previously fixed regions. Roughly speaking, the aim is to transform χ_e to $\mathsf{F}\mathit{core}_e$, to move from ψ_{e-2} to ψ_e. We apply Lévy's zero-one law to deduce that the runs satisfying the χ_e are likely to enter a region that has a high value for a slightly simpler objective, namely

$$\theta_e \stackrel{\text{def}}{=} \varphi_e \wedge \mathsf{G}\,\neg\mathsf{Fix}_{e-2}. \tag{5}$$

To do so, we observe in $\mathcal{L}_{e-2}$ the behavior of several arbitrary $\frac{\gamma}{2}$-optimal strategies σ_{ℓ_0} for ψ_{e-2}, one for each $\ell_0 \in L_0$. Then, for each σ_{ℓ_0}, we apply Lévy's zero-one law separately; this provides that there exists a finite set R_e of states that have a high value for θ_e, and is reached by one of the σ_{ℓ_0} with probability as high as the probability of satisfying the disjunct χ_e. Now we use our previous results [12] on the strategy complexity of Büchi objectives and prove the existence of an MD strategy τ_e that is almost optimal for θ_e (error less than γ), starting from every state in R_e. We define sets fix_e and core_e to be the set of states from which τ_e attains a high probability for θ_e in $\mathcal{L}_{e-2}$; see Figure 3. Define $\beta \stackrel{\text{def}}{=} 1-\gamma$ and $\alpha \stackrel{\text{def}}{=} 1-\gamma^2$, and

$$\mathit{fix}_e \stackrel{\text{def}}{=} \mathit{Safe}_{\mathcal{L}_{e-2},\tau_e,\theta_e}(\beta) \qquad\qquad \mathit{core}_e \stackrel{\text{def}}{=} \mathit{Safe}_{\mathcal{L}_{e-2},\tau_e,\theta_e}(\alpha). \tag{6}$$

We fix the strategy τ_e in the fix_e-region to derive the MDP $\mathcal{L}_e$ from $\mathcal{L}_{e-2}$. Formally,

$$\mathcal{L}_e \stackrel{\text{def}}{=} \mathcal{L}_{e-2}[\tau_e, \mathit{fix}_e]. \tag{7}$$

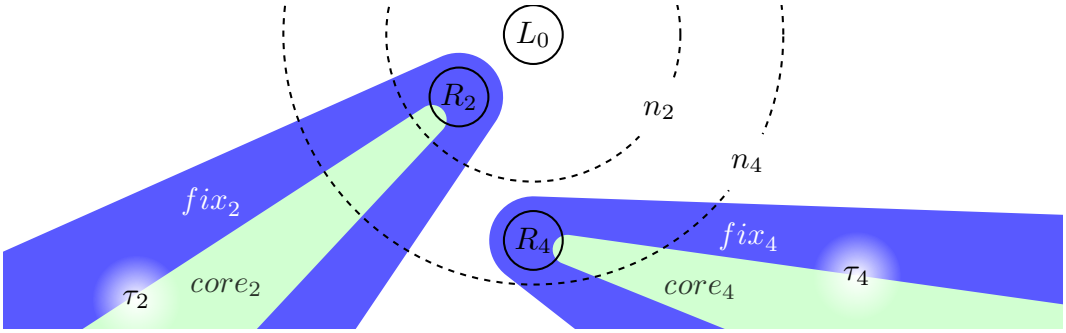


Figure 3 The construction for Lemma 9. In the first iteration, for color 2, we fix the MD strategy τ_2 in the fix_2-region. In the second iteration, for color 4, we fix τ_4 in fix_4, and so on for all even colors. Everywhere else we fix an γ-optimal reachability strategy towards $\bigcup_{e=2}^{e_{\max}} core_e$ (in green).

Iteration $e \in \{2, \cdots, e_{\max}\}$. For all states $\ell_0 \in L_0$, let σ_{ℓ_0} be a general (not necessarily MD) $\frac{\gamma}{2}$-optimal strategy w.r.t. ψ_{e-2} in the layered MDP $\mathcal{L}_{e-2}$. Consider the Markov chain $\mathcal{C}_{\ell_0}$ induced by $\mathcal{L}_{e-2}$, the fixed initial state ℓ_0 and strategy σ_{ℓ_0}.

By definition (Equation 5), θ_e is suffix-closed and $\mathsf{F}\theta_e$ is tail. The strategy σ_{ℓ_0} attains $\mathsf{F}\theta_e$ with probability at least as large as it achieves disjunct χ_e in ψ_{e-2}. We apply Lévy's zero-one law to deduce that the winning runs of $\mathsf{F}\theta_e$ likely reach a finite set R_e of states that have a high value for θ_e. In other words, most runs that eventually win through color e, while eventually avoiding Fix_{e-2}, will reach R_e within a bounded number of steps.

▶ **Lemma 10.** *Let $s_0 \in S$ and $\mathcal{E}$ be a suffix-closed objective. For all $\varepsilon, \varepsilon' > 0$, there exist n and a finite set $F \subseteq Safe_{\mathcal{E}}(1-\varepsilon)$ such that $\mathbb{P}_{s_0}(\mathsf{F}\mathcal{E} \wedge \mathsf{F}^{\leq n} F) \geq \mathbb{P}_{s_0}(\mathsf{F}\mathcal{E}) - \varepsilon'$.*

By Lemma 10, there exist n_{ℓ_0} and a finite set $R_{\ell_0} \subseteq Safe_{\mathcal{L}_{e-2},\theta_e}(\alpha)$ such that

$$\mathbb{P}_{\mathcal{L}_{e-2},\ell_0,\sigma_{\ell_0}}(\mathsf{F}\theta_e \wedge \mathsf{F}^{\leq n_{\ell_0}} R_{\ell_0}) \geq \mathbb{P}_{\mathcal{L}_{e-2},\ell_0,\sigma_{\ell_0}}(\mathsf{F}\theta_e) - \frac{\gamma}{2}. \tag{8}$$

Define $n_e \stackrel{\text{def}}{=} \max_{\ell_0 \in L_0}(n_{\ell_0})$ and $R \stackrel{\text{def}}{=} \bigcup_{\ell_0 \in L_0} R_{\ell_0}$. Write $R_e \stackrel{\text{def}}{=} \{(s,0) \mid \exists b \cdot (s,b) \in R\}$ for the projection of R_e on the layer 0.

▶ Remark 11. Suppose $\mathcal{E}' \subseteq \mathcal{E}$ and $\varepsilon > 0$ are such that $\mathbb{P}(\mathcal{E}') \geq \mathbb{P}(\mathcal{E}) - \varepsilon$. Then, for any $\mathcal{R}$, we have $\mathbb{P}(\mathcal{E}' \cap \mathcal{R}) \geq \mathbb{P}(\mathcal{E} \cap \mathcal{R}) - \varepsilon$.

Proof. We have:

$$\mathbb{P}(\mathcal{E}' \cap \mathcal{R}) = \mathbb{P}(\mathcal{E}') - \mathbb{P}(\mathcal{E}' \setminus \mathcal{R}) \geq \mathbb{P}(\mathcal{E}) - \varepsilon - \mathbb{P}(\mathcal{E}' \setminus \mathcal{R}) \geq \mathbb{P}(\mathcal{E}) - \varepsilon - \mathbb{P}(\mathcal{E} \setminus \mathcal{R}) = \mathbb{P}(\mathcal{E} \cap \mathcal{R}) - \varepsilon.$$

◀

We apply Remark 11 to Equation (8) to get

$$\mathbb{P}_{\mathcal{L}_{e-2},\ell_0,\sigma_{\ell_0}}(\mathsf{F}\theta_e \wedge \mathsf{G}\,\neg\mathsf{Fix}_{e-2} \wedge \mathsf{F}\,Cl(R_e)) \geq \mathbb{P}_{\mathcal{L}_{e-2},\ell_0,\sigma_{\ell_0}}(\mathsf{F}\theta_e \wedge \mathsf{G}\,\neg\mathsf{Fix}_{e-2}) - \frac{\gamma}{2}.$$

Since $\mathsf{FG}\,\neg\mathsf{Fix}_{e-2} \wedge \mathsf{G}\,\neg\mathsf{Fix}_{e-2} = \mathsf{G}\,\neg\mathsf{Fix}_{e-2}$ and $\chi_e = \mathsf{F}\varphi_e \wedge \mathsf{G}\neg\mathsf{Fix}_{e-2}$,

$$\mathbb{P}_{\mathcal{L}_{e-2},\ell_0,\sigma_{\ell_0}}(\chi_e \wedge \mathsf{F}\,Cl(R_e)) \geq \mathbb{P}_{\mathcal{L}_{e-2},\ell_0,\sigma_{\ell_0}}(\chi_e) - \frac{\gamma}{2}. \tag{9}$$

We think of $\mathsf{GF}[S]^{Col=e}$ as a Büchi condition on a slightly modified MDP. This allows us to apply the following theorem from [12] about the strategy complexity of Büchi objectives.

▶ **Theorem 12** (Theorem 5 in [12]). *For every acyclic countable MDP $\mathcal{M}$, a Büchi objective φ, finite set I of initial states and $\varepsilon > 0$, there exists a deterministic 1-bit strategy that is ε-optimal from every $s \in I$.*

Using Theorem 12, we prove the following.

▷ Claim 13. In MDP $\mathcal{L}_{e-2}$, there is an MD strategy τ_e, that is $(\alpha - \beta)$-optimal for θ_e from R_e.

Notice that τ_e is used to define regions $\mathit{core}_e \subseteq \mathit{fix}_e$; see Equation (6) and Figure 3. Since $\mathtt{val}_{\mathcal{L}_{e-2},\theta_e}(\ell) = \mathtt{val}_{\mathcal{L}_{e-2},\theta_e}(\ell')$ holds for all siblings ℓ and ℓ', all states in R_e have value $\geq \alpha$ w.r.t. θ_e. We have chosen τ_e to be $(\alpha - \beta)$-optimal, which implies $\mathbb{P}_{\mathcal{L}_{e-2},\ell,\tau_e}(\theta_e) \geq \beta$ for all $\ell \in R_e$. This shows that $R_e \subseteq \mathit{fix}_e$. Strategy τ_e is also used to obtain $\mathcal{L}_e$ from $\mathcal{L}_{e-2}$: for all controlled states $\ell \in \mathit{fix}_e$, the successor is fixed to be $\tau_e(\ell)$ in $\mathcal{L}_e$, see Equation (7).

Invariant (3)**.** Given a state $\ell_0 \in L_0$, this invariant states that, for all colors e, $\mathtt{val}_{\mathcal{L}_e,\psi_e}(\ell_0) \geq \mathtt{val}_{\mathcal{L},\varphi}(\ell_0) - e \cdot \gamma$ holds. Recall that $\psi_0 = \varphi$ and $\mathcal{L}_0 = \mathcal{L}$. To prove the invariant, by an induction on even colors e, it suffices to prove the following:

$$\mathtt{val}_{\mathcal{L}_e,\psi_e}(\ell_0) \geq \mathtt{val}_{\mathcal{L}_{e-2},\psi_{e-2}}(\ell_0) - 2\gamma.$$

We construct a strategy π for ψ_e in $\mathcal{L}_e$ such that $\mathbb{P}_{\mathcal{L}_e,\ell_0,\pi}(\psi_e) \geq \mathtt{val}_{\mathcal{L}_{e-2},\psi_{e-2}}(\ell_0) - 2\gamma$. Intuitively speaking, π enforces that most runs that win through colors e', with $e' \leq e$, eventually reach the $\mathit{core}_{e'}$-region and most remaining winning runs always avoid the Fix_e-region.

The strategy π is defined by combining σ_{ℓ_0} and τ_e; recall that the strategy σ_{ℓ_0} is $\frac{\gamma}{2}$-optimal w.r.t. ψ_{e-2} starting from ℓ_0 in $\mathcal{L}_{e-2}$. We define π such that it starts by following σ_{ℓ_0}. If it ever enters $\mathit{Cl}(\mathit{fix}_e)$ then we ensure that it enters fix_e as well (in at most one more step). Then π continues by playing as τ_e does forever.

The following claim concludes the proof of **Invariant** (3).

▷ Claim 14. $\mathbb{P}_{\mathcal{L}_e,\ell_0,\pi}(\psi_e) \geq \mathtt{val}_{\mathcal{L}_{e-2},\psi_{e-2}}(\ell_0) - 2\gamma$.

We summarize the main steps in the proof of Claim 14 here. We first prove the claim that if π ever enters $\mathit{Cl}(\mathit{fix}_e)$ then it is possible to define it in such a way that it actually enters fix_e.

Comparing ψ_e with ψ_{e-2}, one notices that two significant terms in the symmetric difference of these two objectives are χ_e and $\mathsf{F}\mathit{core}_e$. Roughly speaking, we use Equation (9) to move from χ_e to $\mathsf{F}\mathit{Cl}(\mathit{fix}_e)$. Then we move from $\mathsf{F}\mathit{Cl}(\mathit{fix}_e)$ to $\mathsf{F}\mathit{core}_e$ by proving that $\mathbb{P}_{\mathcal{L}_e,\ell_0,\pi}(\mathsf{F}\mathit{core}_e)$ is almost as high as $\mathbb{P}_{\mathcal{L}_{e-2},\ell_0,\pi}(\mathsf{F}\mathit{Cl}(\mathit{fix}_e))$, modulo small errors. To derive the latter, we rely on two facts: another application of Lévy's zero-one law that guarantees $\mathbb{P}_{\mathcal{L}_e,\ell_0,\pi}(\theta_e \wedge \mathsf{F}\mathit{core}_e)$ is equal to $\mathbb{P}_{\mathcal{L}_e,\ell_0,\pi}(\theta_e)$; and the fact that, as soon as π visits the first state $\ell \in \mathit{fix}_e$, it switches to τ_e forever, and thus attains θ_e with probability at least β.

Reach iteration. After all $\frac{e_{\max}}{2}$-iterations for even colors and the fixing, by **Invariant** (3), for all $\ell_0 \in L_0$, we have:

$$\mathtt{val}_{\mathcal{L}_{e_{\max}},\psi_{e_{\max}}}(\ell_0) \geq \mathtt{val}_{\mathcal{L},\varphi}(\ell_0) - e_{\max}\gamma. \tag{10}$$

Recall that $\psi_{e_{\max}} = \bigvee_{e \in \mathit{even}(\mathcal{C})} \mathsf{F}\mathit{core}_e$. At this last iteration, we fix the choice of all remaining states in $\mathcal{L}_{e_{\max}}$ such that the probability of $\psi_{e_{\max}}$ is maximized. Recall that there are uniformly ε-optimal MD strategies for reachability objectives [16]. Hence, there is a single MD strategy τ_{reach} in $\mathcal{L}_{e_{\max}}$ that is uniformly γ-optimal w.r.t. $\psi_{e_{\max}}$; in particular, τ_{reach} is γ-optimal from every state $\ell_0 \in L_0$.

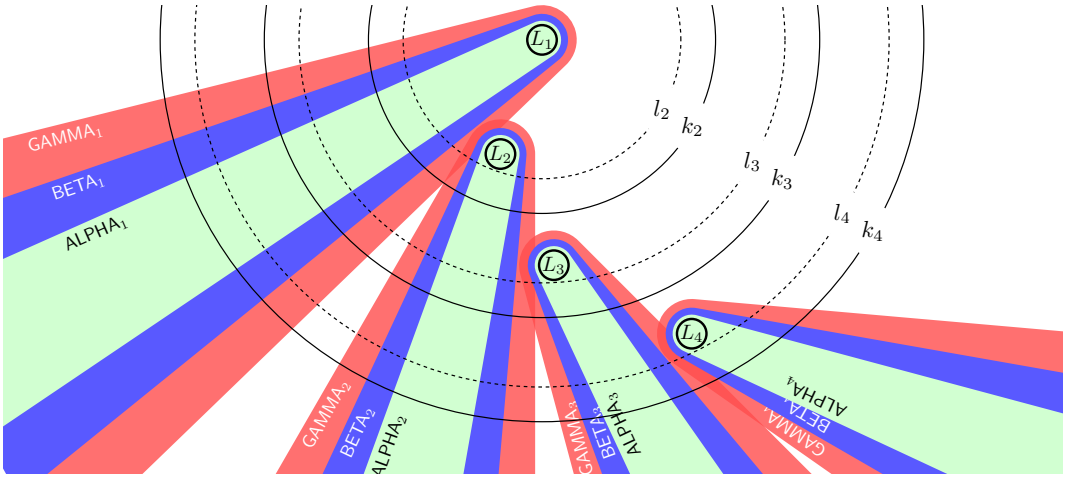


Figure 4 Initial segment of the sea urchin construction. $\mathcal{L}_i$ is the result of fixing τ_i inside BETA_i and then ρ_i inside the k_i-bubble (the set of states reachable from the initial state(s) in $\leq k_i$ steps). Drawn here for $i = 1, 2, 3, 4$.

Let $\mathcal{L}' \stackrel{\text{def}}{=} \mathcal{L}_{e_{\max}}[\tau_{\text{reach}}, L]$. Let $\hat{\sigma}$ be the MD strategy in $\mathcal{L}$ that plays from L_0 as prescribed by all the fixings in $\mathcal{L}'$. Since all choices in all the fix_e-region are resolved according to τ_e, $e \in \{2, \cdots, e_{\max}\}$, we can apply Lévy's zero-one law another time.

▶ **Lemma 15.** *Let $0 < \beta_1 < \beta_2 \leq 1$ and $\mathcal{E}$ a tail objective. For $s \in \mathit{Safe}_{\mathcal{E}}(\beta_2)$, the following holds: $\mathbb{P}_s(\mathsf{G}\,\mathit{Safe}_{\mathcal{E}}(\beta_1)) \geq \frac{\beta_2 - \beta_1}{1 - \beta_1}$.*

By Lemma 15, for all states $\ell \in \mathit{core}_e$,

$$\mathbb{P}_{\mathcal{L}_{e_{\max}},\ell,\tau_e}(\mathsf{G}\mathit{fix}_e) \geq \frac{\alpha - \beta}{1 - \beta} \geq 1 - \gamma. \tag{11}$$

States in fix_e have a high value for θ_e and thus also for $\mathsf{F}\varphi_e$.

▶ **Lemma 16.** *Let $0 < \beta < 1$ and $\mathcal{E}$ a tail objective. For all states $s \in \mathit{Safe}_{\mathcal{E}}(\beta)$:*

1. $\mathbb{P}_s(\mathit{FGSafe}_{\mathcal{E}}(\beta) \setminus \mathcal{E}) = 0$*; and*
2. $\mathbb{P}_s(\mathcal{E} \setminus \mathit{FGSafe}_{\mathcal{E}}(\beta)) = 0$.

By Lemma 16.2, we satisfy $\mathsf{F}\varphi_e$ almost surely:

$$\mathbb{P}_{\mathcal{L}_{e_{\max}},\ell,\tau_e}(\mathsf{F}\varphi_e \mid \mathsf{G}\mathit{fix}_e) = 1. \tag{12}$$

Using Equations (10) and (11), we prove the following.

▷ Claim 17. The MD strategy $\hat{\sigma}$ is ε-optimal for parity objective φ, from every state $\ell_0 \in L_0$.

This concludes the proof of Lemma 9.

4 Optimal Strategies for Parity

In this section we show Theorem 2, i.e., that optimal strategies for parity, where they exist, can be chosen deterministic 1-bit Markov.

First we show the main technical result of this section.

▶ **Lemma 18.** *Let $\mathcal{L}(\mathcal{M})$ be the layered MDP obtained from an acyclic and finitely branching MDP $\mathcal{M}$ and a coloring Col such that all states are almost surely winning for $\varphi = \texttt{Parity}(Col)$ (i.e., every state s has a strategy σ_s such that $\mathbb{P}_{\mathcal{L}(\mathcal{M}),s,\sigma_s}(\varphi) = 1$).*

For every initial state s_0 there exists an MD strategy σ that almost surely wins, i.e., $\mathbb{P}_{\mathcal{L}(\mathcal{M}),s_0,\sigma}(\varphi) = 1$.

Proof sketch. For a complete proof we refer the reader to the technical report [13].

For some intuition consider Figure 4. The sea urchin construction is a plastering construction with infinitely many iterations where MD strategies are fixed in larger and larger subspaces. Its name comes from the shape of the subspace in which player choices are fixed up-to iteration i: A growing finite body of states that are reachable from the initial state s_0 within $\leq k_i$ steps, plus i different spikes of infinite size. Each spike is composed of nested subsets $\mathsf{ALPHA}_i \subseteq \mathsf{BETA}_i$ (and $\subseteq \mathsf{GAMMA}_i$, which is used only in the correctness argument) that correspond to different levels of attainment of certain ε-optimal MD strategies τ_i, obtained from Lemma 9. Strategy τ_i is then fixed in BETA_i (and thus in ALPHA_i). Other MD strategies ρ_i are fixed elsewhere in the finite body, up-to horizon k_i. Using Lévy's zero-one law, we prove that, once inside ALPHA_i, there is a high chance of never leaving the i-th spike BETA_i. Moreover, almost all runs that stay in the i-th spike satisfy parity. Finally, the strategies ρ_i ensure that at least $1/2$ (by probability mass) of the runs from s_0 that don't stay in one of the first i spikes will eventually stay in the $(i+1)$-th spike and satisfy parity there. Thus, at the stage with i spikes, the fixed MD strategy attains parity with some probability $\geq 1 - 2^{-i}$ already *inside* this fixed subspace. In the limit of $i \to \infty$, the resulting MD strategy attains parity almost surely. ◀

▶ **Definition 19.** *For a tail objective φ and an MDP $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P)$, we define the* conditioned version *of $\mathcal{M}$ w.r.t. φ to be the MDP $\mathcal{M}_* = (S_*, S_{*\Box}, S_{*\bigcirc}, \longrightarrow_*, P_*)$ with $S_* = \{s \in S \mid \exists \sigma.\, \mathbb{P}_{\mathcal{M},s,\sigma}(\varphi) = \texttt{val}_{\mathcal{M}}(s) > 0\}$ and $S_{*\Box} = S_* \cap S_\Box$ and $S_{*\bigcirc} = S_* \cap S_\bigcirc$ and*

$$\longrightarrow_* = \{(s,t) \in S_* \times S_* \mid s \longrightarrow t \text{ and if } s \in S_{*\Box} \text{ then } \texttt{val}_{\mathcal{M}}(s) = \texttt{val}_{\mathcal{M}}(t)\}$$

and $P_* : S_{*\bigcirc} \to \mathcal{D}(S_*)$ so that $P_*(s)(t) = P(s)(t) \cdot \frac{\texttt{val}_{\mathcal{M}}(t)}{\texttt{val}_{\mathcal{M}}(s)}$ for all $s \in S_{*\bigcirc}$ and $t \in S_*$ with $s \longrightarrow_* t$.

A proof that $P_*(s)$ is a probability distribution for all $s \in S_{*\bigcirc}$ and therefore that $\mathcal{M}_*$ is well-defined, can be found in the full paper [13], Appendix C. The name "conditional MDP" stems from a useful property that for all strategies that are optimal for φ in $\mathcal{M}$, the probability in $\mathcal{M}_*$ of any event is the same as that of its probability in $\mathcal{M}$ conditioned under φ.

The following theorem is a very slight generalization of [14, Theorem 5]. It gives a sufficient condition under which we can conclude the existence of MD optimal strategies from the existence of MD almost-sure winning strategies.

▶ **Theorem 20.** *Let φ be a tail objective. Let $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P)$ be an MDP and $\mathcal{M}_* = (S_*, S_{*\Box}, S_{*\bigcirc}, \longrightarrow_*, P_*)$ its conditioned version wrt. φ. Then:*

1. *For all $s \in S_*$ there exists a strategy σ with $\mathbb{P}_{\mathcal{M}_*,s,\sigma}(\varphi) = 1$.*
2. *Suppose that for every $s \in S_*$ there exists an MD strategy σ'' with $\mathbb{P}_{\mathcal{M}_*,s,\sigma''}(\varphi) = 1$. Then there is an MD strategy σ' such that for all $s \in S$:*

$$\big(\exists \sigma \in \Sigma.\, \mathbb{P}_{\mathcal{M},s,\sigma}(\varphi) = \texttt{val}_{\mathcal{M}}(s)\big) \implies \mathbb{P}_{\mathcal{M},s,\sigma'}(\varphi) = \texttt{val}_{\mathcal{M}}(s)$$

▶ **Theorem 21.** *Consider an acyclic MDP $\mathcal{M}$ and a parity objective.*

There exists a deterministic 1-bit strategy that is optimal from all states that have an optimal strategy.

Proof. Consider the corresponding layered system $\mathcal{L}(\mathcal{M})$ (cf. Definition 6), which is also acyclic. Let S_{opt} be the subset of states that have an optimal strategy in $\mathcal{M}$. Thus all states in $S_{opt} \times \{0,1\}$ have an optimal strategy in $\mathcal{L}(\mathcal{M})$ by Lemma 7.

We now use Theorem 20 to obtain an MD strategy σ' in $\mathcal{L}(\mathcal{M})$ that is optimal for all states in $\mathcal{L}(\mathcal{M})$ that have an optimal strategy. First, the parity objective is tail. Second, in $\mathcal{L}(\mathcal{M})$, any two siblings have the same value w.r.t. parity by Remark 8. Therefore the changes from $\mathcal{L}(\mathcal{M})$ to its conditioned version $\mathcal{L}(\mathcal{M})_*$ (wrt. the parity objective) are symmetric in the two layers. Thus $\mathcal{L}(\mathcal{M})_*$ is also a layered acyclic MDP (i.e., there exists some acyclic MDP $\mathcal{M}'$ s.t. $\mathcal{L}(\mathcal{M})_* = \mathcal{L}(\mathcal{M}')$), and by Theorem 20.1 all states in $\mathcal{L}(\mathcal{M})_*$ are almost surely winning. Now we can apply Lemma 18 (generalized to infinitely branching acyclic layered MDPs by Lemma 3) to $\mathcal{L}(\mathcal{M})_*$ and obtain that for every state in $\mathcal{L}(\mathcal{M})_*$ there is an MD strategy that almost surely wins. By Theorem 20.2 there is an MD strategy σ' in $\mathcal{L}(\mathcal{M})$ that is optimal for all states that have an optimal strategy. In particular, σ' is optimal for the states in $S_{opt} \times \{0,1\}$ in $\mathcal{L}(\mathcal{M})$. By Lemma 7, this yields a deterministic 1-bit strategy in $\mathcal{M}$ that is optimal for all states in S_{opt}. ◀

In Theorem 21 the initial memory mode of the 1-bit strategy is irrelevant (recall Remark 8).

Theorem 2 now follows directly from Theorem 21 and Lemma 4(3).

5 Optimal Strategies for $\{0,1,2\}$-Parity

▶ **Theorem 22.** *Let $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P)$ be an MDP, φ a $\{0,1,2\}$-Parity objective and $\mathcal{M}_* = (S_*, S_{*\Box}, S_{*\bigcirc}, \longrightarrow_*, P_*)$ its conditioned version wrt. φ. Assume that in $\mathcal{M}_*$ for every safety objective (given by some target $T \subseteq S_*$) and $\varepsilon > 0$ there exists a uniformly ε-optimal MD strategy. Let S_{opt} be the subset of states that have an optimal strategy for φ in $\mathcal{M}$.*

Then there exists an MD strategy in $\mathcal{M}$ that is optimal for φ from every state in S_{opt}.

The above result generalizes [14, Theorem 16], which considers only finitely-branching MDPs and uses the fact that for every safety objective, an MD strategy exists that is uniformly *optimal*. This is not generally true for infinitely-branching acyclic MDPs [14]. To prove Theorem 22, we adjust the construction so that it only requires uniformly ε-*optimal* MD strategies for safety objectives (in the conditioned MDP $\mathcal{M}_*$).

In order to apply Theorem 22 to infinitely-branching acyclic MDPs, we now show that acyclicity guarantees the existence of uniformly ε-optimal MD strategies for safety objectives.

▶ **Lemma 23.** *For every acyclic MDP with a safety objective and every $\varepsilon > 0$ there exists an MD strategy that is uniformly ε-optimal.*

While we defined ε-optimality wrt. additive errors (cf. Section 2), our proof of Lemma 23 shows that the claim holds even wrt. multiplicative errors (in the style of [16]).

▶ **Theorem 24.** *Consider an MDP $\mathcal{M}$ with a $\{0,1,2\}$-Parity objective and let S_{opt} be the subset of states that have an optimal strategy.*

1. *If $\mathcal{M}$ is acyclic then there exists an MD strategy that is optimal from every state in S_{opt}.*
2. *There exists a deterministic Markov strategy that is optimal from every state in S_{opt}.*

Proof. Towards item 1, if $\mathcal{M}$ is acyclic then also its conditioned version $\mathcal{M}_*$ (with respect to $\{0,1,2\}$-Parity) is acyclic. Thus, by Lemma 23, in $\mathcal{M}_*$ for every $\varepsilon > 0$ and every safety objective there is a uniformly ε-optimal MD strategy. The result now follows from Theorem 22.

Item 2 follows from Item 1 and Lemma 4 (item 3 with $k = 0$). ◀

6 ε-Optimal Strategies for $\{0,1\}$-Parity (co-Büchi)

▶ **Theorem 25.** *Suppose that $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P)$ is an MDP such that for every safety objective (given by some target $T \subseteq S$) and $\varepsilon > 0$ there exists a uniformly ε-optimal MD strategy.*

Then for every co-Büchi objective (given by some coloring $Col : S \to \{0,1\}$) and $\varepsilon > 0$ there exists a uniformly ε-optimal MD strategy.

The precondition of Theorem 25 is satisfied by many classes of MDPs. Indeed, we obtain the following.

▶ **Corollary 26.** *Consider an MDP $\mathcal{M}$ and a co-Büchi objective.*

1. *If $\mathcal{M}$ is acyclic then, for every $\varepsilon > 0$, there exists a uniformly ε-optimal MD strategy.*
2. *If $\mathcal{M}$ is finitely branching then, for every $\varepsilon > 0$, there exists a uniformly ε-optimal MD strategy.*
3. *For every $\varepsilon > 0$ there exists a deterministic Markov strategy that, from every initial state s, attains at least $\mathtt{val}_{\mathcal{M}}(s) - \varepsilon$.*

Proof. Towards (1), for acyclic MDPs, uniformly ε-optimal strategies for safety can be chosen MD by Lemma 23. Towards (2), for finitely branching MDPs there always exists even a uniformly optimal MD strategy for every safety objective. In both cases the claim then follows from Theorem 25. Claim (3) follows directly from (1) and Lemma 4 (item 2 with $k = 0$). ◀

References

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

2 C. Baier and J.-P. Katoen. *Principles of Model Checking*. MIT Press, 2008.

3 P. Billingsley. *Probability and Measure*. Wiley, 1995. Third Edition.

4 V. D. Blondel and J. N. Tsitsiklis. A survey of computational complexity results in systems and control. *Automatica*, 2000.

5 N. Bäuerle and U. Rieder. *Markov Decision Processes with Applications to Finance*. Springer-Verlag Berlin Heidelberg, 2011.

6 K. Chatterjee, M. Jurdziński, and T. Henzinger. Quantitative stochastic parity games. In *Annual ACM-SIAM Symposium on Discrete Algorithms*. Society for Industrial and Applied Mathematics, 2004. URL: http://dl.acm.org/citation.cfm?id=982792.982808.

7 E. M. Clarke, T. A. Henzinger, H. Veith, and R. Bloem, editors. *Handbook of Model Checking*. Springer, 2018. doi:10.1007/978-3-319-10575-8.

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

9 E. Grädel, W. Thomas, and T. Wilke, editors. *Automata, Logics, and Infinite Games*, LNCS, 2002.

10 T. P. Hill. On the existence of good Markov strategies. *Transactions of the American Mathematical Society*, 1979. doi:10.1090/S0002-9947-1979-0517690-9.

11 T. P. Hill. Goal problems in gambling theory. *Revista de Matemática: Teoría y Aplicaciones*, 1999.

12 S. Kiefer, R. Mayr, M. Shirmohammadi, and P. Totzke. Büchi objectives in countable MDPs. In *International Colloquium on Automata, Languages and Programming*. LIPIcs, 2019. A technical report is available at arXiv:1904.11573. doi:10.4230/LIPIcs.ICALP.2019.119.

13 S. Kiefer, R. Mayr, M. Shirmohammadi, and P. Totzke. Strategy Complexity of Parity Objectives in Countable MDPs. *CoRR*, 2020. arXiv:2007.05065.

14 S. Kiefer, R. Mayr, M. Shirmohammadi, and D. Wojtczak. Parity objectives in countable MDPs. In *Annual IEEE Symposium on Logic in Computer Science*, 2017.

15 A. Mostowski. Regular expressions for infinite trees and a standard form of automata. In *Computation Theory*, LNCS, 1984.

16 D. Ornstein. On the existence of stationary optimal strategies. *Proceedings of the American Mathematical Society*, 1969.

17 M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1st edition, 1994.

18 M. Schäl. Markov decision processes in finance and dynamic options. In *Handbook of Markov Decision Processes*. Springer, 2002.

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

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

21 M.Y. Vardi. Automatic verification of probabilistic concurrent finite-state programs. In *Proc. of FOCS'85*, 1985.

22 W. Zielonka. Infinite games on finitely coloured graphs with applications to automata on infinite trees. *Theoretical Computer Science*, 1998.

23 W. Zielonka. Perfect-information stochastic parity games. In *Foundations of Software Science and Computation Structures*, LNCS. Springer, 2004.