Büchi Objectives in Countable MDPs

Citation for published version:
Kiefer, S, Mayr, R, Shirmohammadi, M & Totzke, P 2019, Büchi Objectives in Countable MDPs. in C Baier, I Chatzigiannakis, P Flocchini & S Leonardi (eds), *46th International Colloquium on Automata, Languages, and Programming (ICALP 2019).* vol. 132, 119, LIPICS, vol. 132, Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany, pp. 119:1–119:14, 46th International Colloquium on Automata, Languages and Programming, Patras, Greece, 8/07/19. https://doi.org/10.4230/LIPIcs.ICALP.2019.119

Digital Object Identifier (DOI):
10.4230/LIPIcs.ICALP.2019.119

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
46th International Colloquium on Automata, Languages, and Programming (ICALP 2019)

Büchi Objectives in Countable MDPs

Stefan Kiefer
University of Oxford, UK

Richard Mayr
University of Edinburgh, UK

Mahsa Shirmohammadi
CNRS, Paris, France
IRIF, Paris, France

Patrick Totzke
University of Liverpool, UK

Abstract

We study countably infinite Markov decision processes with Büchi objectives, which ask to visit a given subset F of states infinitely often. A question left open by T.P. Hill in 1979 [10] is whether there always exist ε-optimal Markov strategies, i.e., strategies that base decisions only on the current state and the number of steps taken so far. We provide a negative answer to this question by constructing a non-trivial counterexample. On the other hand, we show that Markov strategies with only 1 bit of extra memory are sufficient.

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

Keywords and phrases Markov decision processes

Digital Object Identifier 10.4230/LIPIcs.ICALP.2019.119

Category Track B: Automata, Logic, Semantics, and Theory of Programming

Related Version The full version of this paper is available at `https://arxiv.org/abs/1904.11573`.

Funding *Stefan Kiefer*: Supported by a Royal Society University Research Fellowship.
Richard Mayr: Supported by EPSRC grant EP/M027651/1.
Mahsa Shirmohammadi: Supported by PEPS JCJC grant AAPS.

Acknowledgements The authors thank anonymous reviewers for their helpful comments.

1 Introduction

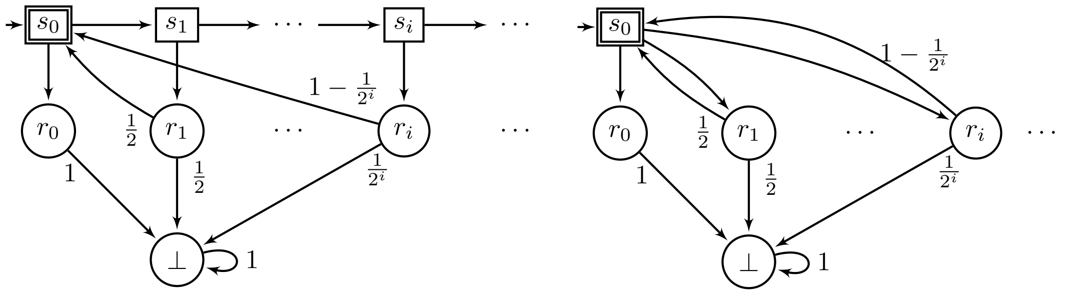
Background. Markov decision processes (MDPs) are a standard model for dynamic systems that exhibit both stochastic and controlled behavior [16]. MDPs play a prominent role in numerous domains, including artificial intelligence and machine learning [19, 18], control theory [4, 1], operations research and finance [5, 17], and formal verification [9, 2]. In an MDP, the system starts in the initial state and makes a sequence of transitions between states. Depending on the type of the current state, either the controller gets to choose an enabled transition (or a distribution over transitions), or the next transition is chosen randomly according to a defined distribution. By fixing a strategy for the controller, 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 needed for an optimal (resp. ε-optimal) strategy for some objective is also called the *strategy complexity* of the objective. 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;

46th International Colloquium on Automata, Languages, and Programming (ICALP 2019).
Editors: Christel Baier, Ioannis Chatzigiannakis, Paola Flocchini, and Stefano Leonardi;
Article No. 119; pp. 119:1–119:14
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

(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 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 or the run. Thus they do use infinite memory, but only in a very restricted form by maintaining an unbounded step-counter. For finite MDPs, there exist optimal MD-strategies for many (but not all) objectives [6, 7, 8, 16], but the picture is more complex for countably infinite MDPs [13, 15, 16].

We study here so-called *Goal* objectives defined via a subset of goal states F: In the basic Goal objective (also called the *Reachability* objective) one simply wants to reach the set F. In the *Büchi* objective one wants to visit the set F infinitely often. For finite MDPs there exist optimal MD-strategies for both these objectives [7, 16]. For countably infinite MDPs, optimal strategies (where they exist) and ε-optimal strategies for Reachability can be chosen MD [15, 16]. Similarly, optimal strategies for Büchi (where they exist) can be chosen MD [13]. However, ε-optimal strategies for Büchi require infinite memory (cannot be chosen FR); cf. [13, 14].



(a) Finitely branching, but infinitely many controlled states.

(b) Infinitely branching, but just one controlled state.

Figure 1 Two MDPs where ε-optimal strategies for Büchi require infinite memory. Let $F = \{s_0\}$ be the set of goal states. Here and throughout the paper we indicate goal states by double borders, and controlled states as rectangles.

▶ **Example 1.** Consider the MDPs in Figure 1. Every finite memory (FR) strategy will only attain probability 0 for Büchi in these examples [13]. However, there exists an ε-optimal Markov strategy for every $\varepsilon > 0$: At the i-th time that state s_0 is visited, pick the successor state r_{i+k} where k is some sufficiently large number depending on ε, e.g., $k = \lceil \log_2(1/\varepsilon) \rceil$. For example (b) this can easily be done with a step-counter since s_0 is visited for the i-th time in step $2(i-1)$ unless the system has reached the state $\perp$. For example (a), under this strategy, state s_0 is visited for the i-th time in step $\sum_{j=1}^{i-1}(k+j+1)$ unless the system has reached the state $\perp$. ◀

▶ **Example 2.** Consider the MDP from Figure 2, taken from [11, Example 4.2]. Every FR-strategy attains only probability 0 of Büchi. Moreover, the strategy that, in state s_0, subsequently picks $r_1, r_2, \ldots$ also attains probability 0, unlike in Example 1. But a different infinite-memory strategy achieves a positive probability. Indeed, let σ be the strategy that, in s_0, picks 2^1 times r_1 and then 2^2 times r_2 and $\ldots 2^i$ times r_i etc. This strategy σ achieves a positive probability of Büchi. (In more detail, σ achieves a positive probability of not falling in a losing sink $\perp$, and in almost all of the remaining runs it visits a goal state infinitely often.) Note that σ is a Markov strategy. ◀

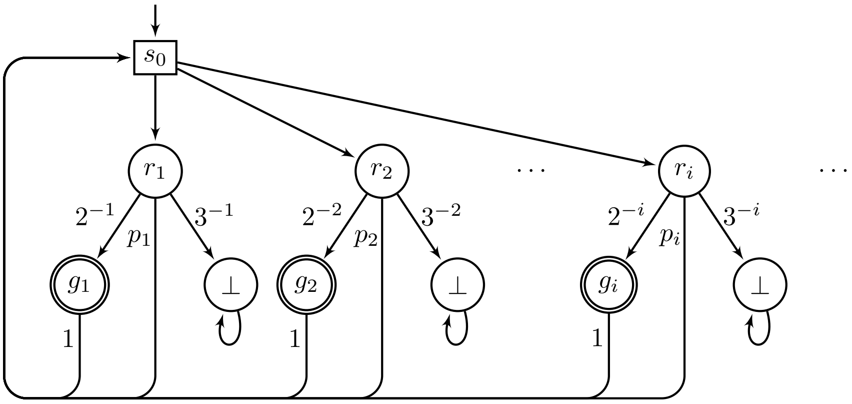


Figure 2 An MDP where ε-optimal strategies for Büchi require infinite memory. The transition probability p_i stands for $1 - 2^{-i} - 3^{-i}$. The state s_0 is the only controlled state.

The open problem. While the MDPs in Examples 1 and 2 require infinite memory, Markov strategies suffice for them. Such examples led to the question whether there always exists a family of ε-optimal Markov strategies for Büchi in all countably infinite MDPs.

A partial answer was given by Hill [10] (Proposition 5.1), who showed that ε-optimal Markov strategies for Büchi exist in the special case where the MDP contains only a *finite* number of controlled states. This result applies to the MDPs from Example 2 and Figure 1b), but not directly to the one in Figure 1a).

The question for general MDPs was stated as an open problem in [10] (p.158, l.4) and mentioned again in [11] (Q1 in Section 5).

Our contributions. We provide a negative answer to the open problem. We construct a non-trivial example of a countable acyclic and finitely branching MDP and prove that no ε-optimal Markov strategies for Büchi exist for it (for any $\varepsilon < 1$). In combination with the example from Figure 1, this shows that for general MDPs neither finite memory (FR) nor Markov strategies are sufficient.

Secondly, we provide an upper bound on the strategy complexity of Büchi. We show that for *acyclic* countable MDPs there always exist ε-optimal strategies that are deterministic and use only one bit of memory. Since every MDP can be transformed into an acyclic one by encoding a step-counter into the states, it follows that general countable MDPs have ε-optimal strategies for Büchi that are deterministic and use only a step-counter plus one extra bit of memory. Thus Markov strategies are almost, but not quite, sufficient. Table 1 summarizes these results.

Table 1 Existence of various types of ε-optimal strategies for the Büchi objective, for several classes of MDPs. New results are in boldface.

ε-optimal strategy for Büchi	MD	1-bit D	FR	Markov	Markov+1 bit D
Finite MDP	✓	✓	✓	✓	✓
MDP w. finitely many controlled states	×	×	×	✓	✓
Acyclic MDP	**×**	**✓**	**✓**	**×**	**✓**
General MDP	×	×	×	**×**	**✓**

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.

For a set S we write S^* (resp. S^ω) for the set of all finite (resp. infinite) sequences of elements in S. We use slightly generalized regular expressions for sets of sequences, e.g., if $s_0 \in S$ we may write $s_0 S^\omega$ for the set of infinite sequences starting with s_0.

Markov decision processes. A *Markov decision process* (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 (non-empty) set of successor states. A *sink in* $\mathcal{M}$ is a subset $T \subseteq S$ closed under the $\longrightarrow$ relation, that is, $s \in T$ and $s{\longrightarrow}s'$ implies that $s' \in T$.

An MDP is *acyclic* if the underlying directed graph $(S, {\longrightarrow})$ is acyclic, i.e., there is no directed cycle. It is *finitely branching* if every state has finitely many successors and *infinitely branching* otherwise. An MDP without controlled states ($S_\Box = \emptyset$) is called a *Markov chain.*

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 $\rho(i) \overset{\text{def}}{=} s_i$ 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 $\{s' \in S \mid s{\longrightarrow}s'\}$. The set of all strategies in $\mathcal{M}$ is denoted by $\Sigma_\mathcal{M}$ (we omit the subscript and write Σ if $\mathcal{M}$ is clear from the context). A (partial) run $s_0 s_1 \cdots$ 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 $P(s_i)(s_{i+1}) > 0$.

An MDP $\mathcal{M} = (S, S_\Box, S_\bigcirc, {\longrightarrow}, P)$, an initial state $s_0 \in S$, and a strategy σ induce a probability space in which the outcomes are runs starting in s_0 and with measure $\mathcal{P}_{\mathcal{M},s_0,\sigma}$ defined as follows. It is first defined on *cylinders* $s_0 s_1 \ldots s_n S^\omega$, where $s_1, \ldots, s_n \in S$: if $s_0 s_1 \ldots s_n$ is not a partial run induced by σ then $\mathcal{P}_{\mathcal{M},s_0,\sigma}(s_0 s_1 \ldots s_n S^\omega) \overset{\text{def}}{=} 0$. Otherwise, $\mathcal{P}_{\mathcal{M},s_0,\sigma}(s_0 s_1 \ldots s_n S^\omega) \overset{\text{def}}{=} \prod_{i=0}^{n-1} \bar{\sigma}(s_0 s_1 \ldots s_i)(s_{i+1})$, 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 $\mathcal{P}_{\mathcal{M},s_0,\sigma}$ on the Borel σ-algebra $\mathcal{F}$ of subsets of $s_0 S^\omega$. Elements of $\mathcal{F}$, i.e., measurable sets of runs, are called *events* or *objectives* here. For $X \in \mathcal{F}$ we will write $\overline{X} \overset{\text{def}}{=} s_0 S^\omega \setminus X \in \mathcal{F}$ for its complement and $\mathcal{E}_{\mathcal{M},s_0,\sigma}$ for the expectation w.r.t. $\mathcal{P}_{\mathcal{M},s_0,\sigma}$. We drop the indices wherever possible without introducing ambiguity.

Strategy Classes. Strategies are in general *randomized* (R) in the sense that they take values in $\mathcal{D}(S)$. A strategy σ is *deterministic* (D) if $\sigma(\rho)$ is a Dirac distribution for all 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, and let $\tau : \mathsf{M} \times S \to \mathcal{D}(\mathsf{M} \times S)$ be a function that meets the following two conditions: for all modes $\mathsf{m} \in \mathsf{M}$,

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

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

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

- A strategy σ is *finite memory* (F) if there exists a finite memory M implementing σ.
- 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 *1-bit* if it can be implemented with a memory of size 2. Such a strategy is then determined by a function $\tau : \{0, 1\} \times S \to \mathcal{D}(\{0, 1\} \times S)$. Intuitively τ uses one bit of memory to capture two different modes.
- A strategy σ is *Markov* if it can be implemented with the natural numbers $\mathbb{N}$ as the memory, and a function τ such that the distribution $\tau(\mathsf{m}, s)$ 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 the current state and the number of steps taken so far, i.e., it has access to a step-counter. We view Markov strategies as functions $\sigma : \mathbb{N} \times S_\Box \to \mathcal{D}(S)$. Note that such a strategy is generally not finite memory.
- A strategy σ is *1-bit Markov* if it can be implemented with $\mathbb{N} \times \{0, 1\}$ as the memory, and a function τ such that the distribution $\tau(n, b, s)$ is over $\{n+1\} \times \{0, 1\} \times S$ for all $(n, b) \in \mathsf{M}$ and $s \in S$. We view such strategies as functions $\sigma : \mathbb{N} \times \{0, 1\} \times S_\Box \to \mathcal{D}(\{0, 1\} \times S)$.

Payoffs, Values, Optimality. We are interested in strategies to maximize the expectation of a given measurable *payoff* function $f : S^\omega \to \mathbb{R}$, a random variable that assigns a real value to every run. The *value* of state s (w.r.t. f) is the supremum of expected values of f over all strategies:

$$\mathtt{val}_{\mathcal{M},f}(s) \overset{\text{def}}{=} \sup_{\sigma \in \Sigma} \mathcal{E}_{\mathcal{M},s,\sigma}(f),$$

For $\varepsilon \geq 0$ and $s \in S$, we say that a strategy σ is *ε-optimal* iff $\mathcal{E}_{\mathcal{M},s,\sigma}(f) \geq \mathtt{val}_{\mathcal{M},f}(s) - \varepsilon$ and *uniformly* ε-optimal iff this holds for every $s \in S$. A (uniformly) 0-optimal strategy is simply called (uniformly) *optimal*.

In this paper, we will need two types of payoff functions. The first is the *total reward*, a random variable given as $f(\rho) \overset{\text{def}}{=} \sum_{t=0}^{\infty} r(\rho(t))$, where $r : S \to \mathbb{R}$ is some given *reward* function. A useful fact [16, Theorem 7.1.9] is that if S is finite and the range of r is bounded then there exist optimal strategies (for total reward) which are memoryless and deterministic.

The second type of payoff functions we consider are those with range $\{0, 1\}$. Each such payoff function f uniquely identifies an objective (set of runs) φ by viewing f as the characteristic function of φ, i.e., $f(\rho) = 1$ if $\rho \in \varphi$ and 0 otherwise. Then $\mathcal{E}_{\mathcal{M},s,\sigma}(f) = \mathcal{P}_{\mathcal{M},s,\sigma}(\varphi)$. We call this the *probability of achieving* φ (using strategy σ starting from the state s) and simply write $\mathtt{val}_{\mathcal{M},\varphi}(s) = \mathtt{val}_{\mathcal{M},f}(s) = \sup_{\sigma \in \Sigma} \mathcal{P}_{\mathcal{M},s,\sigma}(\varphi)$.

Our main focus are *reachability* (sometimes also called *goal*) and *Büchi* objectives, which are determined by a set of states $F \subseteq S$ and defined as follows. Let us slightly abuse notation and identify F with its characteristic function, i.e., $F(s) = 1$ if $s \in F$.

- The *reachability* objective is to visit F at least once during a run. The corresponding payoff is $f(\rho) \overset{\text{def}}{=} \max_{t \in \mathbb{N}} \rho(t)$, and we define $\mathtt{Goal}(F) \overset{\text{def}}{=} \{\rho \in S^\omega \mid \max_{t \in \mathbb{N}} F(\rho(t)) = 1\}$;
- The *Büchi* objective is to visit F infinitely often. The corresponding payoff function is $f(\rho) \overset{\text{def}}{=} \limsup_{t \to \infty} F(\rho(t))$, and we let $\mathtt{Büchi}(F) \overset{\text{def}}{=} \{\rho \in S^\omega \mid \limsup_{t \to \infty} F(\rho(t)) = 1\}$.

3 The Lower Bound

In this section we solve Hill's problem ([10] and [11, Q1]) by exhibiting an MDP where the initial state has value 1 w.r.t. the Büchi objective, but every Markov strategy achieves this objective with probability 0. As explained in the introduction, it follows that in acyclic MDPs, ε-optimal MR-strategies are not guaranteed to exist. In fact, in the following theorem we prove the latter fact first, and subsequently generalize it to solve Hill's problem.

▶ **Theorem 3.** *There exists an acyclic MDP $\mathcal{M}$, a state s_0 and a set of states F such that*

1. *for every Markov strategy σ, we have $\mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathtt{Büchi}(F)) = 0$, and*
2. $\mathtt{val}_{\mathtt{Büchi}(F)}(s_0) = 1$ *and for every $\varepsilon > 0$ there exists a deterministic 1-bit strategy σ_ε s.t. $\mathcal{P}_{\mathcal{M},s_0,\sigma_\varepsilon}(\mathtt{Büchi}(F)) \geq 1 - \varepsilon$.*

In the remainder of this section we provide a proof sketch. The full proof is in [12].

Proof sketch for Theorem 3. Our construction is based on an infinite MDP $\mathcal{M}$ that consists of a chain of height-n trees, T^n, for $n \in \mathbb{N} = \{1, 2, \ldots\}$. Figure 3 depicts its initial segment T^1, T^2, T^3. Each such tree is "rooted" at a brown state on the top level, with a transition incoming from a blue state. We make use of some conventions that simplify the presentation and the analysis. In Figure 3, the different colors of the states highlight the structure of the MDP; the colors are also indicated by letters in the states: blue (L), brown (B), yellow (Y), red (R), green (G), white (W). The start state, s_0, is the blue state in the top-left corner. The controlled states are exactly the yellow states. The goal set F consists of the green states at the bottom. Two transitions emanate from each red state: a black (right) transition and a red (left) transition, both leading to the same (brown or green) state.

We consider the strengthened Büchi objective that asks to see F infinitely often and moreover that *no red transition* is taken. This corresponds exactly to the normal Büchi objective if we redirect every red transition to an infinite (losing) chain of non-green states (not depicted in Figure 3).

We first argue that no MR-strategy achieves a positive probability of that objective. Then we show that the MDP $\mathcal{M}$ can be modified so that no Markov strategy achieves a positive probability.

Intuition behind the construction of $\mathcal{M}$. The objective, say φ, of visiting infinitely many green states and no red transition creates tension between trying to visit green states and avoiding too many red states (the latter states incur a risk of taking a red transition). In the proof we need to show that no memoryless strategy strikes a good balance between these competing goals. On the one end of the spectrum, an MR-strategy might always choose the upward transition in the yellow states (which are the only controlled states). But such a strategy never visits any green state, thus clearly violates φ. On the other end of the spectrum lies the "greedy" MR-strategy, which always chooses the downward transition in the yellow states, in order to visit as many green states as possible. Indeed, under this strategy, let u_n denote the probability that, starting in the top-left brown state of T^n, no green state is visited in T^n. By induction (given in [12]) one can show that there is $u < 1$ such that $u_n \leq u$ holds for all n. Considering the probability of the transitions emanating from the blue states (at the top), the expected overall number of visited green states is at least $\sum_{n=1}^{\infty} \frac{1}{n}(1 - u_n) \geq (1 - u) \sum_{n=1}^{\infty} \frac{1}{n} = \infty$. It is not hard to strengthen this statement so that the greedy strategy almost surely visits infinitely many green states. So the greedy strategy satisfies one part of φ, but it does so at the expense of visiting many red states. Red states though are associated with a risk of taking a red transition, and it follows from the proof in [12] that the greedy strategy almost surely ends up taking at least one (and indeed infinitely many) red transition(s).

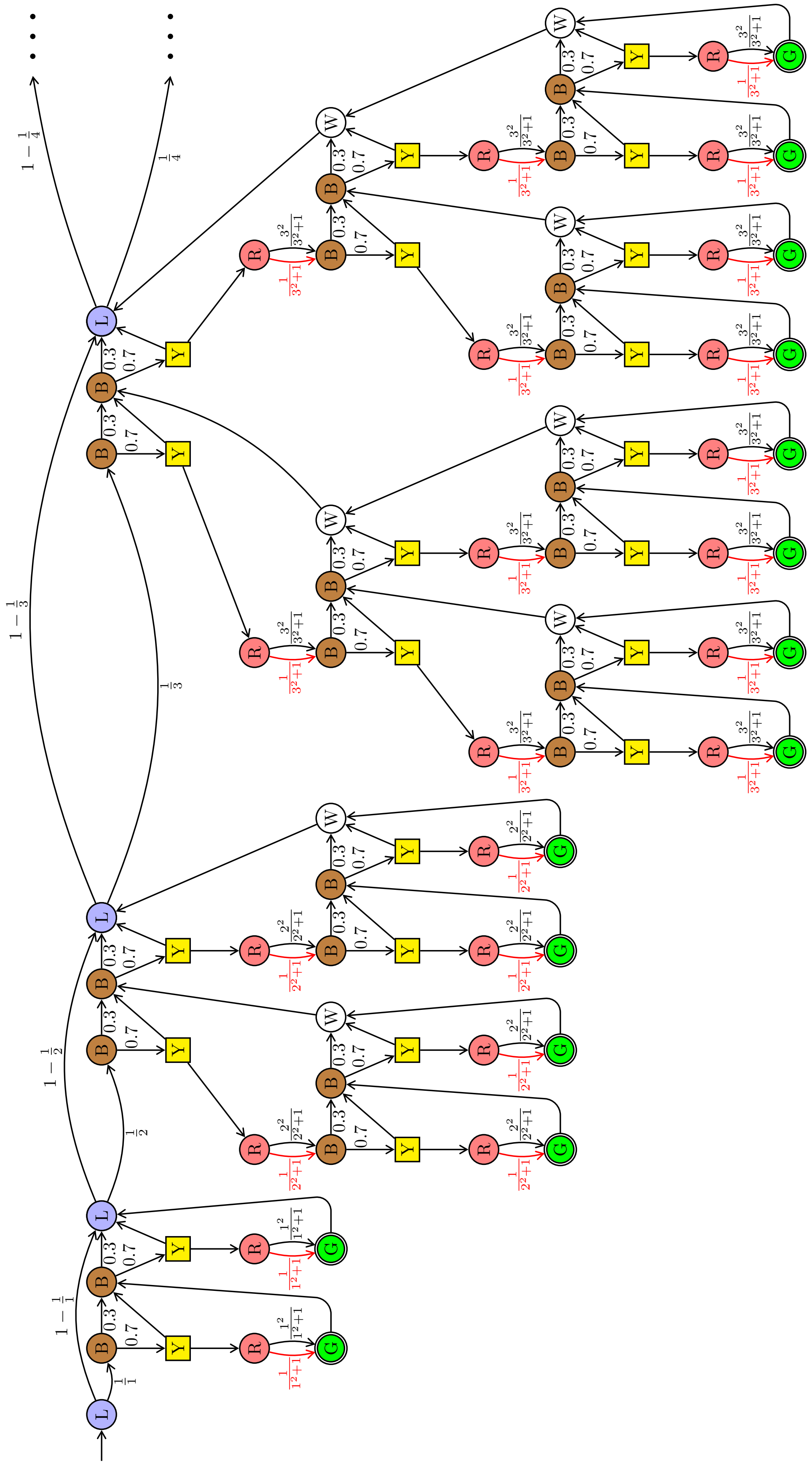


Figure 3 For this acyclic MDP $\mathcal{M}$ there are ε-optimal deterministic 1-bit strategies for the Büchi objective $\texttt{Büchi}(T)$ where T contains exactly all green states. No MR-strategy achieves even a positive probability.

Good 1-bit strategies. The two competing goals discussed in the previous paragraph can be balanced using a deterministic 1-bit strategy, which we describe in the following. This strategy, σ_1, sets its bit to 0 whenever a blue state (at the top) is entered. While the bit is 0, in each tree T^n it maximizes the probability of visiting a green state by choosing the downward transition in the yellow states, thus accepting a certain risk of taking a red transition. However, if and when a green state in T^n is visited, the bit is set to 1, and for the remaining sojourn in T^n the strategy σ_1 chooses the upward transitions in the yellow states, thus avoiding any risk of a red transition in the remainder of T^n. Although σ_1 appears to visit fewer green states than the aforementioned "greedy" MR-strategy, σ_1 still visits infinitely many green states almost surely. This is because for each tree T^n, the two strategies have the same probability of visiting at least one green state in T^n. The strategy σ_1 can be improved, for each $\varepsilon > 0$, to achieve φ with probability at least $1 - \varepsilon$, by fixing the bit to 1 in the first k trees $T^1, \ldots, T^k$, for a k that depends on ε. Thus the first k trees are virtually skipped, eliminating the risk of taking any red transition there. In this way one can make the risk of taking a red transition arbitrarily small, while still visiting infinitely many green states with probability 1.

No good MR-strategies. We need to show that not only the extreme MR-strategies described above are inadequate but that every MR-strategy achieves φ with probability 0. To this end, for each tree T^n, define two probabilities:

- t_n (for "total success"): the probability that, starting in the top-left brown state of T^n, at least one green state but no red transition is visited in T^n;
- d_n (for "death"): the probability that, starting in the top-left brown state of T^n, a red transition is visited in T^n.

A very technical proof shows that $d_n \geq 0.008 \cdot t_n$ holds for all n, and this key inequality captures the inability of *any* MR-strategy to strike an adequate balance between the mentioned competing goals. Indeed, one can show that for an MR-strategy to have a positive probability of not visiting any red transition, the series $\sum_{n=1}^{\infty} \frac{1}{n} \cdot d_n$ needs to converge; but to have a positive probability of visiting infinitely many green states, the series $\sum_{n=1}^{\infty} \frac{1}{n} \cdot t_n$ needs to diverge (in both cases, the factor $\frac{1}{n}$ is the probability of visiting the top-left brown node of T^n). By the inequality above, this is impossible.

No good Markov strategies. For the proof of Theorem 3, we also need to show that all Markov strategies achieve probability 0. To this end, we modify the MDP $\mathcal{M}$ so that for each state, all paths from the initial state s_0 to s have the same length. This can be achieved by replacing some transitions in $\mathcal{M}$ by longer chains consisting of non-green states. This modification does not change the fact that MR-strategies achieve probability 0. But since in the new MDP each state can only be visited at a certain time, which is known a priori, a step-counter does not help. Hence all Markov strategies, like MR-strategies, achieve φ with probability 0. ◀

Theorem 3 answers Hill's question negatively. By combining the MDP from Theorem 3 with one of the MDPs from Figure 1 (by adding a new initial random state that branches to the MDPs with probability $\frac{1}{2}$ each), one can even construct a single MDP whose value w.r.t. $\texttt{Büchi}(F)$ is 1, but every FR- and every Markov strategy achieves probability 0.

A slight modification of the example above yields a lower bound on the memory requirements for the almost-sure parity objective. Recall that the parity objective is defined on systems whose states are labeled by a finite set of colors $C \overset{\text{def}}{=} \{1, 2, \ldots, \mathit{max}\} \subseteq \mathbb{N}$, where a run is in $\texttt{Parity}(C)$ iff the highest color that is seen infinitely often in the run is even.

▶ **Corollary 4.** *There exist an acyclic MDP $\mathcal{M}'$ with colors $\{1,2,3\}$ and a state s_0 such that*

1. *for every Markov strategy σ, we have $\mathcal{P}_{\mathcal{M}',s_0,\sigma}(\mathtt{Parity}(\{1,2,3\})) = 0$, and*
2. *there exists a deterministic 1-bit strategy σ' such that $\mathcal{P}_{\mathcal{M}',s_0,\sigma'}(\mathtt{Parity}(\{1,2,3\})) = 1$.*

Proof. We obtain $\mathcal{M}'$ by modifying the MDP $\mathcal{M}$ from Theorem 3 as follows. Label all green states in F by color 2 and the rest by color 1. Then modify each red transition to go to its target via a fresh state labeled by color 3. Clearly $\mathcal{M}'$ is still acyclic and labeled by colors $\{1,2,3\}$.

From the proof of Theorem 3 (1), under every Markov strategy in $\mathcal{M}$ a.s. seeing infinitely many green states (in F) implies seeing infinitely many red transitions. So in $\mathcal{M}'$ every Markov strategy σ a.s. either sees color 2 only finitely often or color 3 infinitely often, thus $\mathcal{P}_{\mathcal{M}',s_0,\sigma}(\mathtt{Parity}(\{1,2,3\})) = 0$.

From the proof of Theorem 3 (2), there is a deterministic 1-bit strategy σ in $\mathcal{M}$ that attains probability $\geq 1/2$ for $\mathtt{Büchi}(F)$ without taking any red transition and otherwise a.s. takes a red transition. This property of σ holds not only when starting from s_0 but from every other state as well. We obtain σ' in $\mathcal{M}'$ by continuing to play σ even after red transitions have been taken. Under σ' the probability of going through infinitely many red transitions (and seeing color 3) is $\leq (1/2)^\infty = 0$, and the probability of seeing infinitely many states in F (with color 2) is 1. Thus $\mathcal{P}_{\mathcal{M}',s_0,\sigma'}(\mathtt{Parity}(\{1,2,3\})) = 1$. ◀

4 The Upper Bound

We show that acyclic MDPs admit ε-optimal deterministic 1-bit strategies for Büchi.

We start by giving some intuition why 1 bit of memory is needed and how it is used. A step $s' \longrightarrow s''$ from some controlled state s' is *value-decreasing* iff $\mathtt{val}(s'') < \mathtt{val}(s')$. While an optimal strategy can never tolerate any value-decreasing step, an ε-optimal strategy might have to take value-decreasing steps infinitely often. The trick is to keep the collective value-loss sufficiently small ($\leq \varepsilon$), while satisfying the other requirements of the objective. So the strategy needs to play “ever better” (i.e., tolerate only smaller and smaller value decreases) along a run. In general this requires infinite memory, since one might re-visit the same state infinitely often and needs to choose a different transition from it every time; cf. Figure 1. However, in an acyclic MDP, with high probability, the distance to the initial state increases with the number of steps taken. Thus one can partition the state space into separate regions, depending on the distance from the initial state, and fix an acceptable rate of value-decrease for each region. Just limiting the collective value-loss is not sufficient for Büchi, one also needs to make progress and visit the set of goal states F at least once in each region. The problem is that some runs might linger in some region too long, and visit F *many times*, but see too many value-decreasing steps at the rate of this region. Therefore, as soon as one has visited F in some region, one should try to get to the next outer region (further away from the initial state) where the rate of value-loss is smaller. Thus one needs 1 bit of memory to record whether one has already seen F in this region. (Remember that the same state can be reached by different runs with different histories.) Just 1 bit suffices, because the probability of returning to a previous inner region (and misinterpreting the bit) can be made arbitrarily small, since the MDP is acyclic.

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

Proof. Let $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P)$ be an acyclic MDP, $I \subseteq S$ a finite set of initial states and $F \subseteq S$ a set of goal states and $\varphi \overset{\text{def}}{=} \texttt{Büchi}(F)$ denote the Büchi objective w.r.t. F. We prove the claim for finitely branching $\mathcal{M}$ first and transfer the result to general MDPs at the end. For every $\varepsilon > 0$ and every $s \in I$ there exists an ε-optimal strategy σ_s such that

$$\mathcal{P}_{\mathcal{M},s,\sigma_s}(\varphi) \geq \texttt{val}_{\mathcal{M},\varphi}(s) - \varepsilon. \tag{1}$$

However, the strategies σ_s might differ from each other and might use randomization and a large (or even infinite) amount of memory. We will construct a single deterministic strategy σ' that uses only 1 bit of memory such that $\forall_{s\in I}\, \mathcal{P}_{\mathcal{M},s,\sigma'}(\varphi) \geq \texttt{val}_{\mathcal{M},\varphi}(s) - 2\varepsilon$. This proves the claim as ε can be chosen arbitrarily small.

In order to construct σ', we first observe the behavior of the finitely many σ_s for $s \in I$ on an infinite, increasing sequence of finite subsets of S. Based on this, we define a second stronger objective φ' with

$$\varphi' \subseteq \varphi, \tag{2}$$

and show that all σ_s attain at least $\texttt{val}_{\mathcal{M},\varphi}(s) - 2\varepsilon$ w.r.t. φ', i.e.,

$$\forall_{s\in I}\, \mathcal{P}_{\mathcal{M},s,\sigma_s}(\varphi') \geq \texttt{val}_{\mathcal{M},\varphi}(s) - 2\varepsilon. \tag{3}$$

We construct σ' as a deterministic 1-bit *optimal* strategy w.r.t. φ' from all $s \in I$ and obtain

$$\begin{array}{rcll} \mathcal{P}_{\mathcal{M},s,\sigma'}(\varphi) & \geq & \mathcal{P}_{\mathcal{M},s,\sigma'}(\varphi') & \text{by (2)} \\ & \geq & \mathcal{P}_{\mathcal{M},s,\sigma_s}(\varphi') & \text{by optimality of } \sigma' \text{ for } \varphi' \\ & \geq & \texttt{val}_{\mathcal{M},\varphi}(s) - 2\varepsilon & \text{by (3).} \end{array}$$

Informal outline: Behavior of σ_s, objective φ' and properties (2) and (3). For the formal proof see [12].

Let $\mathsf{bubble}_k(I)$ be the set of states that can be reached from some initial state in I within at most k steps. Since I is finite and $\mathcal{M}$ is finitely branching, $\mathsf{bubble}_k(I)$ is finite for every k.

We define a sequence of sufficiently large and increasing numbers k_i and l_i with $k_i < l_i < k_{i+1}$ for $i \in \mathbb{N}$ and finite sets $K_i \overset{\text{def}}{=} \mathsf{bubble}_{k_i}(I)$ and $L_i \overset{\text{def}}{=} \mathsf{bubble}_{l_i}(I)$. Every run from a $s \in I$ according to σ_s must eventually leave each of these finite sets, because $\mathcal{M}$ is acyclic. Moreover, we choose these numbers so that once a run has left L_i it is very unlikely to return to K_i. Let $F_i \overset{\text{def}}{=} F \cap K_i \setminus L_{i-1}$. Runs according to σ_s are very likely to follow a particular pattern. Let $R_1 \overset{\text{def}}{=} (K_1 \setminus F_1)^* F_1$, $R_2 \overset{\text{def}}{=} (K_2 \setminus F_2)^* F_2$ and $R_{i+1} \overset{\text{def}}{=} (K_{i+1} \setminus (F_{i+1} \cup K_{i-1}))^* F_{i+1}$ for $i \geq 2$. We show that

$$\forall_{s\in I}\, \mathcal{P}_{\mathcal{M},s,\sigma_s}(\varphi \cap \overline{R_1 R_2 \dots R_{i+1} (S \setminus K_i)^\omega}) \quad \leq \quad \varepsilon \tag{4}$$

We now define the Borel objectives $R_{\leq i} \overset{\text{def}}{=} R_1 R_2 \dots R_i S^\omega$ and $\varphi' \overset{\text{def}}{=} \bigcap_{i\in\mathbb{N}} R_{\leq i}$. Since $F_i \cap F_k = \emptyset$ for $i \neq k$ and φ' implies a visit to the set F_i for all $i \in \mathbb{N}$, we have $\varphi' \subseteq \varphi$ and obtain (2). Using (4), we show that $\forall_{s\in I}\, \mathcal{P}_{\mathcal{M},s,\sigma_s}(\varphi') \geq \texttt{val}_{\mathcal{M},\varphi}(s) - 2\varepsilon$ and thus obtain (3).

Definition of the 1-bit strategy σ'. We now define a deterministic 1-bit strategy σ' that is optimal for objective φ' from every $s \in I$. First we define certain "suffix" objectives of φ'. Recall that $R_i = (K_i \setminus (F_i \cup K_{i-2}))^* F_i$. Let $R_{i,j} \overset{\text{def}}{=} R_i R_{i+1} \dots R_j S^\omega$ and $R_{\geq i} \overset{\text{def}}{=} \bigcap_{j\geq i} R_{i,j}$. Consider the objectives $R_{\geq i+1}$ for runs that start in states $s' \in F_i$. For every state $s' \in F_i$ we consider its value w.r.t. the objective $R_{\geq i+1}$, i.e., $\texttt{val}_{\mathcal{M},R_{\geq i+1}}(s') \overset{\text{def}}{=} \sup_{\hat{\sigma}} \mathcal{P}_{\mathcal{M},s',\hat{\sigma}}(R_{\geq i+1})$.

For every $i \geq 1$ we consider the finite subspace $K_i \setminus K_{i-2}$. In particular, it contains the sets F_{i-1} and F_i. We define a bounded total reward objective B_i for runs starting in F_{i-1} as follows. Runs that exit the subspace (either by leaving K_i or by visiting K_{i-2}) before visiting F_i get reward 0. All other runs must visit F_i eventually (since $\mathcal{M}$ is acyclic and the subspace is finite). When some run reaches the set F_i *for the first time* in some state s' then this run gets the reward of $\texttt{val}_{\mathcal{M},R_{\geq i+1}}(s')$. Using [16, Theorem 7.1.9], we show that there exists a uniform optimal MD-strategy σ_i for B_i on $K_i \setminus K_{i-2}$ in $\mathcal{M}$.

We now define σ' by combining different MD-strategies σ_i, depending on the current state and on the value of the 1-bit memory. The intuition is that the strategy σ' has two modes: normal-mode and next-mode. In a state $s' \in K_i \setminus K_{i-1}$, if the memory is $i \pmod 2$ then the strategy is in normal-mode and plays towards reaching F_i. Otherwise, the strategy is in next-mode and plays towards reaching F_{i+1}.

Initially σ' starts in a state $s \in I$ with the 1-bit memory set to 1. We define the behavior of σ' in a state $s' \in K_i \setminus K_{i-1}$ for every $i \geq 1$. If the 1-bit memory is $i \pmod 2$ and $s' \notin F_i$ then σ' plays like σ_i. (Intuitively, one plays towards F_i, since one has not yet visited it.) If the 1-bit memory is $i \pmod 2$ and $s' \in F_i$ then the 1-bit memory is set to $(i+1) \pmod 2$, and σ' plays like σ_{i+1}. (Intuitively, one records the fact that one has already seen F_i and then targets the next set F_{i+1}.) If the 1-bit memory is $(i+1) \pmod 2$ then σ' plays like σ_{i+1}. (Intuitively, one plays towards F_{i+1}, since one has already visited F_i.)

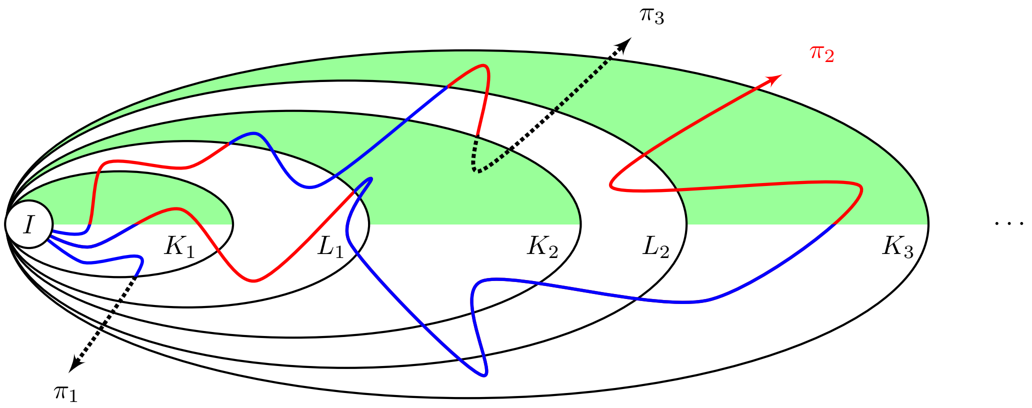


Figure 4 Memory updates along runs π_1, π_2, π_3, drawn in blue while the memory-bit is one and in red while the bit is zero. The green region in K_1 is F_1, and for all $i \geq 2$, the green region in $K_i \setminus L_{i-1}$ is F_i. Both π_1 and π_3 violate φ' and are drawn as dotted lines once they do.

Observe that if a run according to σ' exits some set K_i (and thus enters $K_{i+1} \setminus K_i$) with the bit still set to $i \pmod 2$ (normal-mode) then this run has not visited F_i and thus does not satisfy the objective φ'. (Or the same has happened earlier for some $j < i$, in which case also the objective φ' is violated.) An example is the run π_1 in Figure 4. However, if a run according to σ' exits some set K_i (and thus enters $K_{i+1} \setminus K_i$) with the bit set to $(i+1) \pmod 2$ (thus σ_{i+1} in next-mode) then in the new set $K_{i'} \setminus K_{i'-1}$ with $i' = i+1$ the bit is set to $i' \pmod 2$ and σ' continues to play like σ_{i+1} in normal-mode. Even if this run returns (temporarily) to K_i (but not to K_{i-1}) the strategy σ' continues to play like σ_{i+1} in next-mode. An example is the run π_2 in Figure 4. Finally, if a run returns to K_{i-1} after having visited F_i then it fails the objective φ', e.g., run π_3 in Figure 4.

The 1-bit strategy σ' is optimal for φ' from every $s \in I$. Let $s \in I$ be arbitrary. For a given run from s, let $\mathsf{firstin}(F_i)$ be the first state s' in F_i that is visited (if any). We define a bounded reward objective B_i' for runs starting at s as follows. Every run that does not satisfy the objective $R_{\leq i}$ gets assigned reward 0. Otherwise, consider a run from s that satisfies $R_{\leq i}$. When this run reaches the set F_i for the first time in some state s' then this run gets a reward of $\mathtt{val}_{\mathcal{M},R_{\geq i+1}}(s')$. Note that this reward is ≤ 1.

We show that for all $i \in \mathbb{N}$

$$\mathtt{val}_{\mathcal{M},\varphi'}(s) = \mathtt{val}_{\mathcal{M},B_i'}(s) \tag{5}$$

Towards the $\geq$ inequality, let $\hat{\sigma}$ be an $\hat{\varepsilon}$-optimal strategy for B_i' from s. We define the strategy $\hat{\sigma}'$ to play like $\hat{\sigma}$ until a state $s' \in F_i$ is reached and then to switch to some $\hat{\varepsilon}$-optimal strategy for objective $R_{\geq i+1}$ from s'. Every run from s that satisfies φ' can be split into parts, before and after the first visit to the set F_i, i.e., $\varphi' = \{w_1 s' w_2 \mid w_1 s' \in R_{\leq i}, s' \in F_i, s' w_2 \in R_{\geq i+1}\}$. Therefore we obtain that $\mathcal{P}_{\mathcal{M},s,\hat{\sigma}'}(\varphi') \geq \mathcal{E}_{\mathcal{M},s,\hat{\sigma}}(B_i') - \hat{\varepsilon} \geq \mathtt{val}_{\mathcal{M},B_i'}(s) - 2\hat{\varepsilon}$. Since this holds for every $\hat{\varepsilon} > 0$, we obtain $\mathtt{val}_{\mathcal{M},\varphi'}(s) \geq \mathtt{val}_{\mathcal{M},B_i'}(s)$.

Towards the $\leq$ inequality, let $\hat{\sigma}$ be any strategy for φ' from s. We have $\mathcal{P}_{\mathcal{M},s,\hat{\sigma}}(\varphi') \leq \sum_{s' \in F_i} \mathcal{P}_{\mathcal{M},s,\hat{\sigma}}(R_{\leq i} \cap \mathsf{firstin}(F_i) = s') \cdot \mathtt{val}_{\mathcal{M},R_{\geq i+1}}(s') = \mathcal{E}_{\mathcal{M},s,\hat{\sigma}}(B_i')$. Thus $\mathtt{val}_{\mathcal{M},\varphi'}(s) \leq \mathtt{val}_{\mathcal{M},B_i'}(s)$. Together we obtain (5).

For all $i \in \mathbb{N}$ and every state $s' \in F_i$ we show that

$$\mathtt{val}_{\mathcal{M},R_{\geq i+1}}(s') = \mathtt{val}_{\mathcal{M},B_{i+1}}(s') \tag{6}$$

Towards the $\geq$ inequality, let $\hat{\sigma}$ be an $\hat{\varepsilon}$-optimal strategy for B_{i+1} from $s' \in F_i$. We define the strategy $\hat{\sigma}'$ to play like $\hat{\sigma}$ until a state $s'' \in F_{i+1}$ is reached and then to switch to some $\hat{\varepsilon}$-optimal strategy for objective $R_{\geq i+2}$ from s''. We have that $\mathcal{P}_{\mathcal{M},s',\hat{\sigma}'}(R_{\geq i+1}) \geq \mathcal{E}_{\mathcal{M},s',\hat{\sigma}}(B_{i+1}) - \hat{\varepsilon} \geq \mathtt{val}_{\mathcal{M},B_{i+1}}(s) - 2\hat{\varepsilon}$. Since this holds for every $\hat{\varepsilon} > 0$, we obtain $\mathtt{val}_{\mathcal{M},R_{\geq i+1}}(s') \geq \mathtt{val}_{\mathcal{M},B_{i+1}}(s')$.

Towards the $\leq$ inequality, let $\hat{\sigma}$ be any strategy for $R_{\geq i+1}$ from $s' \in F_i$. We have

$$\begin{aligned}\mathcal{P}_{\mathcal{M},s',\hat{\sigma}}(R_{\geq i+1}) \;&\leq \sum_{s'' \in F_{i+1}} \mathcal{P}_{\mathcal{M},s',\hat{\sigma}}(R_{i+1} S^\omega \cap \mathsf{firstin}(F_{i+1}) = s'') \cdot \mathtt{val}_{\mathcal{M},R_{\geq i+2}}(s'')\\ &= \mathcal{E}_{\mathcal{M},s',\hat{\sigma}}(B_{i+1}).\end{aligned}$$

Thus $\mathtt{val}_{\mathcal{M},R_{\geq i+1}}(s') \leq \mathtt{val}_{\mathcal{M},B_{i+1}}(s')$. Together we obtain (6).

We show, by induction on i, that σ' is optimal for B_i' for all $i \in \mathbb{N}$ from start state s, i.e.,

$$\mathcal{E}_{\mathcal{M},s,\sigma'}(B_i') = \mathtt{val}_{\mathcal{M},B_i'}(s) \tag{7}$$

In the base case of $i = 1$ we have that $B_1' = B_1$. The strategy σ' plays σ_1 until reaching F_1, which is optimal for objective B_1 and thus optimal for B_1'. For the induction step we assume (IH) that σ' is optimal for B_i'.

$$\begin{aligned}
\mathtt{val}_{\mathcal{M},B_{i+1}'}(s) \;&=\; \mathtt{val}_{\mathcal{M},B_i'}(s) && \text{by (5)}\\
&=\; \mathcal{E}_{\mathcal{M},s,\sigma'}(B_i') && \text{by (IH)}\\
&=\; \sum_{s' \in F_i} \mathcal{P}_{\mathcal{M},s,\sigma'}(R_{\leq i} \cap \mathsf{firstin}(F_i) = s') \cdot \mathtt{val}_{\mathcal{M},R_{\geq i+1}}(s') && \text{by def. of } B_i'\\
&=\; \sum_{s' \in F_i} \mathcal{P}_{\mathcal{M},s,\sigma'}(R_{\leq i} \cap \mathsf{firstin}(F_i) = s') \cdot \mathtt{val}_{\mathcal{M},B_{i+1}}(s') && \text{by (6)}\\
&=\; \sum_{s' \in F_i} \mathcal{P}_{\mathcal{M},s,\sigma'}(R_{\leq i} \cap \mathsf{firstin}(F_i) = s') \cdot \mathcal{E}_{\mathcal{M},s',\sigma_{i+1}}(B_{i+1}) && \text{opt. of } \sigma_{i+1} \text{ for } B_{i+1}\\
&=\; \mathcal{E}_{\mathcal{M},s,\sigma'}(B_{i+1}') && \text{by def. of } \sigma' \text{ and } B_{i+1}'
\end{aligned}$$

So σ' attains the value $\mathtt{val}_{\mathcal{M},B'_{i+1}}(s)$ of the objective B'_{i+1} from s and is optimal. Thus (7).

Now we show that σ' performs well on the objectives $R_{\leq i}$ for all $i \in \mathbb{N}$.

$$\mathcal{P}_{\mathcal{M},s,\sigma'}(R_{\leq i}) \geq \mathtt{val}_{\mathcal{M},\varphi'}(s) \tag{8}$$

We have

$$\begin{aligned}
\mathcal{P}_{\mathcal{M},s,\sigma'}(R_{\leq i}) &\geq \mathcal{E}_{\mathcal{M},s,\sigma'}(B'_i) && \text{since } B'_i \text{ gives rewards 0 for runs} \notin R_{\leq i} \text{ and} \leq 1 \text{ otherwise} \\
&= \mathtt{val}_{\mathcal{M},B'_i}(s) && \text{by (7)} \\
&= \mathtt{val}_{\mathcal{M},\varphi'}(s) && \text{by (5)}
\end{aligned}$$

So we get (8). Now we are ready to prove the optimality of σ' for φ' from s.

$$\begin{aligned}
\mathcal{P}_{\mathcal{M},s,\sigma'}(\varphi') &= \mathcal{P}_{\mathcal{M},s,\sigma'}(\cap_{i\in\mathbb{N}} R_{\leq i}) && \text{by def. of } \varphi' \\
&= \lim_{i\to\infty} \mathcal{P}_{\mathcal{M},s,\sigma'}(R_{\leq i}) && \text{by continuity of measures from above} \\
&\geq \lim_{i\to\infty} \mathtt{val}_{\mathcal{M},\varphi'}(s) && \text{by (8)} \\
&= \mathtt{val}_{\mathcal{M},\varphi'}(s)
\end{aligned}$$

From finitely to infinitely branching MDPs. Encode infinite branching into finite branching like in Figure 1, apply the above result to obtain a 1-bit strategy for the finitely branching version, and then transform this strategy back into a 1-bit strategy for the original MDP. ◀

Now we show our upper bound on the strategy complexity of Büchi for general MDPs.

▶ **Theorem 6.** *For every countable MDP $\mathcal{M}$, finite set of initial states I, set of states F and $\varepsilon > 0$, there exists a deterministic 1-bit Markov strategy for $\mathtt{B\ddot{u}chi}(F)$ that is ε-optimal from every $s \in I$.*

Proof. Encode a step-counter into the states to obtain an acyclic MDP, apply Theorem 5 to obtain an ε-optimal deterministic 1-bit strategy for it, and then transform this strategy back into an ε-optimal deterministic 1-bit Markov strategy in the original MDP. ◀

References

1 Pieter Abbeel and Andrew Y. Ng. Learning first-order Markov models for control. In *Advances in Neural Information Processing Systems 17*, pages 1–8. MIT Press, 2004.

2 Christel Baier and Joost-Pieter Katoen. *Principles of Model Checking*. MIT Press, 2008.

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

4 Vincent D. Blondel and John N. Tsitsiklis. A survey of computational complexity results in systems and control. *Automatica*, 36(9):1249–1274, 2000.

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

6 K. Chatterjee, L. de Alfaro, and T. Henzinger. Trading memory for randomness. In *Annual Conference on Quantitative Evaluation of Systems*, pages 206–217. IEEE Computer Society Press, 2004.

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

8 K. Chatterjee, M. Jurdziński, and T. Henzinger. Quantitative Stochastic Parity Games. In *Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 121–130. Society for Industrial and Applied Mathematics, 2004.

9 Edmund M. Clarke, Thomas A. Henzinger, Helmut Veith, and Roderick Bloem, editors. *Handbook of Model Checking*. Springer, 2018.

10 Theodore Preston Hill. On the Existence of Good Markov strategies. *Transactions of the American Mathematical Society*, 247:157–176, 1979.

11 Theodore Preston Hill. Goal Problems in Gambling Theory. *Revista de Matemática: Teoría y Aplicaciones*, 6(2):125–132, 1999.

12 Stefan Kiefer, Richard Mayr, Mahsa Shirmohammadi, and Patrick Totzke. Büchi Objectives in Countable MDPs. Technical report, arxiv.org, 2019. `arXiv:1904.11573`.

13 Stefan Kiefer, Richard Mayr, Mahsa Shirmohammadi, and Dominik Wojtczak. Parity Objectives in Countable MDPs. In *Annual IEEE Symposium on Logic in Computer Science*, 2017.

14 J. Krčál. Determinacy and Optimal Strategies in Stochastic Games. Master's thesis, Masaryk University, School of Informatics, 2009.

15 D. Ornstein. On the existence of stationary optimal strategies. *Proceedings of the American Mathematical Society*, 20:563–569, 1969.

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

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

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

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