Transience in Countable MDPs

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

Richard Mayr
School of Informatics, University of Edinburgh, UK

Mahsa Shirmohammadi
Université de Paris, CNRS, IRIF, F-75013 Paris, France

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

Abstract

The `Transience` objective is not to visit any state infinitely often. While this is not possible in any finite Markov Decision Process (MDP), it can be satisfied in countably infinite ones, e.g., if the transition graph is acyclic.

We prove the following fundamental properties of `Transience` in countably infinite MDPs.

1. There exist uniformly ε-optimal MD strategies (memoryless deterministic) for `Transience`, even in infinitely branching MDPs.
2. Optimal strategies for `Transience` need not exist, even if the MDP is finitely branching. However, if an optimal strategy exists then there is also an optimal MD strategy.
3. If an MDP is universally transient (i.e., almost surely transient under all strategies) then many other objectives have a lower strategy complexity than in general MDPs. E.g., ε-optimal strategies for Safety and co-Büchi and optimal strategies for $\{0,1,2\}$-`Parity` (where they exist) can be chosen MD, even if the MDP is infinitely branching.

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, Transience

Digital Object Identifier 10.4230/LIPIcs.CONCUR.2021.11

Related Version *Full Version*: `https://arxiv.org/abs/2012.13739`

1 Introduction

> Those who cannot remember the past are condemned to repeat it.
>
> *George Santayana (1905) [22]*

The famous aphorism above has often been cited (with small variations), e.g., by Winston Churchill in a 1948 speech to the House of Commons, and carved into several monuments all over the world [22].

We prove that the aphorism is false. In fact, even those who cannot remember anything at all are *not* condemned to repeat the past. With the right strategy they can avoid repeating the past equally well as everyone else. More formally, playing for `Transience` does not require any memory. We show that there always exist ε-optimal memoryless deterministic strategies for `Transience`, and if optimal strategies exist then there also exist optimal memoryless deterministic strategies.[1]

[1] Our result applies to MDPs (also called games against nature). It is an open question whether it generalizes to countable stochastic 2-player games. (However, it is easy to see that the adversary needs infinite memory in general, even if the player is passive [14, 16].)

Background. We study Markov decision processes (MDPs), a standard model for dynamic systems that exhibit both stochastic and controlled behavior [21]. MDPs play a prominent role in many domains, e.g., artificial intelligence and machine learning [26, 24], control theory [5, 1], operations research and finance [25, 12, 6, 23], and formal verification [2, 25, 11, 8, 3, 7].

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 *strategy complexity* of a given objective characterizes the type of strategy necessary to achieve an optimal (resp. ε-optimal) value for the objective. General strategies can take the whole history of the run into account (history-dependent; (H)), while others use only bounded information about it (finite memory; (F)) or base decisions 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.

Acyclicity and Transience. An MDP is called acyclic iff its transition graph is acyclic. While finite MDPs cannot be acyclic (unless they have deadlocks), countable MDPs can. In acyclic countable MDPs, the strategy complexity of Büchi/Parity objectives is lower than in the general case: ε-optimal strategies for Büchi/Parity objectives require only one bit of memory in acyclic MDPs, while they require infinite memory (an unbounded step-counter, plus one bit) in general countable MDPs [14, 15].

The concept of *transience* can be seen as a generalization of acyclicity. In a Markov chain, a state s is called *transient* iff the probability of returning from s to s is < 1 (otherwise the state is called recurrent). This means that a transient state is almost surely visited only finitely often. The concept of transient/recurrent is naturally lifted from Markov chains to MDPs, where they depend on the chosen strategy.

We define the `Transience` objective as the set of runs that do not visit any state infinitely often. We call an MDP *universally transient* iff it almost-surely satisfies `Transience` under every strategy. Thus every acyclic MDP is universally transient, but not vice-versa; cf. Figure 1. In particular, universal transience does not just depend on the structure of the transition graph, but also on the transition probabilities. Universally transient MDPs have interesting properties. Many objectives (e.g., Safety, Büchi, co-Büchi) have a lower strategy complexity than in general MDPs; see below.

We also study the strategy complexity of the `Transience` objective itself, and how it interacts with other objectives, e.g., how to attain a Büchi objective in a transient way.

Our contributions.

1. We show that there exist uniformly ε-optimal MD strategies (memoryless deterministic) for `Transience`, even in infinitely branching MDPs. This is unusual, since (apart from reachability objectives) most other objectives require infinite memory if the MDP is infinitely branching, e.g., all objectives generalizing Safety [17].

   Our result is shown in several steps. First we show that there exist ε-optimal deterministic 1-bit strategies for `Transience`. Then we show how to dispense with the 1-bit memory and obtain ε-optimal MD strategies for `Transience`. Finally, we make these MD strategies uniform, i.e., independent of the start state.

2. We show that optimal strategies for `Transience` need not exist, even if the MDP is finitely branching. If they do exist then there are also MD optimal strategies. More generally, there exists a single MD strategy that is optimal from every state that allows optimal strategies for `Transience`.
3. If an MDP is universally transient (i.e., almost surely transient under all strategies) then many other objectives have a lower strategy complexity than in general MDPs, e.g., ε-optimal strategies for Safety and co-Büchi and optimal strategies for $\{0,1,2\}$-`Parity` (where they exist) can be chosen MD, even if the MDP is infinitely branching.

For our proofs we develop some technical results that are of independent interest. We generalize Ornstein's plastering construction [20] from reachability to tail objectives and thus obtain a general tool to infer uniformly ε-optimal MD strategies from non-uniform ones (cf. Theorem 7). Secondly, in Section 6 we develop the notion of the *conditioned MDP* (cf. [17]). For tail objectives, this allows to obtain uniformly ε-optimal MD strategies wrt. *multiplicative errors* from those with merely additive errors.

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.

Markov Decision Processes. We define Markov decision processes (MDPs for short) over countably infinite state spaces as tuples $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P)$ where S is the countable set of states partitioned into a set $S_\Box$ of *controlled states* and a set $S_\bigcirc$ of *random states*. The *transition relation* is $\longrightarrow \subseteq S \times S$, and $P : S_\bigcirc \to \mathcal{D}(S)$ is a *probability function*. 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 $i \in \mathbb{N}$; a *partial run* is a finite prefix of a run. We write $\rho(i) = s_i$ and say that (partial) run $s_0 s_1 \cdots$ *visits* s if $s = s_i$ for some i. It *starts in* s if $s = s_0$.

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. We write $\Sigma_\mathcal{M}$ for the set of all strategies in $\mathcal{M}$. A strategy σ and an initial state $s_0 \in S$ induce a standard probability measure on sets of infinite runs. We write $\mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathfrak{R})$ for the probability of a measurable set $\mathfrak{R} \subseteq s_0 S^\omega$ of runs starting from s_0. It is defined for the cylinders $s_0 s_1 \dots s_n S^\omega \in S^\omega$ as $\mathcal{P}_{\mathcal{M},s_0,\sigma}(s_0 s_1 \dots s_n S^\omega) \stackrel{\text{def}}{=} \prod_{i=0}^{n-1} \bar{\sigma}(s_0 s_1 \dots s_i)(s_{i+1})$, 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 [4], the measure for cylinders extends uniquely to a probability measure $\mathcal{P}_{\mathcal{M},s_0,\sigma}$ on all measurable subsets of $s_0 S^\omega$. We will write $\mathcal{E}_{\mathcal{M},s_0,\sigma}$ for the expectation w.r.t. $\mathcal{P}_{\mathcal{M},s_0,\sigma}$.

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 $\rho \in S^* S_\Box$.

A formal definition of the amount of *memory* needed to implement strategies can be found in the full version [13]. The two classes of *memoryless* and *1-bit* strategies are central to this paper. A strategy σ is *memoryless* (M) if σ bases its decision only on the last state of the run: $\sigma(\rho s) = \sigma(\rho' s)$ for all $\rho, \rho' \in S^*$. We may view M-strategies as functions $\sigma : S_\Box \to \mathcal{D}(S)$. A 1-bit strategy σ may base its decision also on a memory mode $\mathsf{m} \in \{0,1\}$. Formally, a 1-bit strategy σ is given as a tuple (u, m_0) where $\mathsf{m}_0 \in \{0,1\}$ is the initial memory mode and $u : \{0,1\} \times S \to \mathcal{D}(\{0,1\} \times S)$ is an update function such that

- for all controlled states $s \in S_\Box$, the distribution $u((\mathsf{m}, s))$ is over $\{0,1\} \times \{s' \mid s \longrightarrow s'\}$.
- for all random states $s \in S_\bigcirc$, we have that $\sum_{\mathsf{m}' \in \{0,1\}} u((\mathsf{m}, s))(\mathsf{m}', s') = P(s)(s')$.

Note that this definition allows for updating the memory mode upon visiting random states. We write $\sigma[\mathsf{m}_0]$ for the strategy obtained from σ by setting the initial memory mode to m_0.

MD strategies are both memoryless and deterministic; and *deterministic 1-bit* strategies are both deterministic and 1-bit.

Objectives. The objective of the controller is determined by a predicate on infinite runs. We assume familiarity with the syntax and semantics of the temporal logic LTL [9]. Formulas are interpreted on the underlying structure $(S, \longrightarrow)$ of the MDP $\mathcal{M}$. We use $[\![\varphi]\!]^{\mathcal{M},s} \subseteq sS^\omega$ to denote the set of runs starting from s that satisfy the LTL formula φ, which is a measurable set [27]. We also write $[\![\varphi]\!]^{\mathcal{M}}$ for $\bigcup_{s \in S} [\![\varphi]\!]^{\mathcal{M},s}$. Where it does not cause confusion we will identify φ and $[\![\varphi]\!]$ and just write $\mathcal{P}_{\mathcal{M},s,\sigma}(\varphi)$ instead of $\mathcal{P}_{\mathcal{M},s,\sigma}([\![\varphi]\!]^{\mathcal{M},s})$.

Given a set $T \subseteq S$ of states, the *reachability* objective $\texttt{Reach}(T) \stackrel{\text{def}}{=} \mathsf{F}T$ is the set of runs that visit T at least once. The *safety* objective $\texttt{Safety}(T) \stackrel{\text{def}}{=} \mathsf{G}\neg T$ 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

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

We write $\mathcal{C}$-$\texttt{Parity}$ for the parity objectives with the set of colors $\mathcal{C} \subseteq \mathbb{N}$. The classical Büchi and co-Büchi objectives correspond to $\{1,2\}$-$\texttt{Parity}$ and $\{0,1\}$-$\texttt{Parity}$, respectively.

An objective φ is called a *tail objective* (in $\mathcal{M}$) iff for every run $\rho'\rho$ with some finite prefix ρ' we have $\rho'\rho \in \varphi \Leftrightarrow \rho \in \varphi$. For every coloring Col, $\texttt{Parity}(Col)$ is tail. Reachability objectives are not always tail but in MDPs where the target set T is a sink $\texttt{Reach}(T)$ is tail.

Optimal and ε-optimal Strategies. Given an objective φ, the *value* of state s in an MDP $\mathcal{M}$, denoted by $\texttt{val}_{\mathcal{M},\varphi}(s)$, is the supremum probability of achieving φ. Formally, we have $\texttt{val}_{\mathcal{M},\varphi}(s) \stackrel{\text{def}}{=} \sup_{\sigma \in \Sigma} \mathcal{P}_{\mathcal{M},s,\sigma}(\varphi)$ 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 $\mathcal{P}_{\mathcal{M},s,\sigma}(\varphi) \geq \texttt{val}_{\mathcal{M},\varphi}(s) - \varepsilon$. A 0-optimal strategy is called *optimal*. An optimal strategy is *almost-surely winning* iff $\texttt{val}_{\mathcal{M},\varphi}(s) = 1$.

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 $s \in S$.

Throughout the paper, we may drop the subscripts and superscripts from notations, if it is understood from the context. The missing proofs can be found in the full version [13].

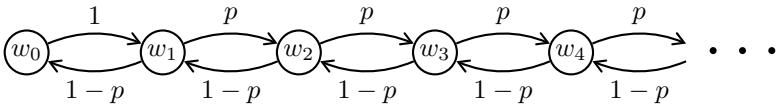

Figure 1 Gambler's Ruin with restart: The state w_i illustrates that the controller's wealth is i, and the coin tosses are in the controller's favor with probability p. For all i, $\mathcal{P}_{w_i}(\texttt{Transience}) = 0$ if $p \leq \frac{1}{2}$; and $\mathcal{P}_{w_i}(\texttt{Transience}) = 1$ otherwise.

3 Transience and Universally Transient MDPs

In this section we define the transience property for MDPs, a natural generalization of the well-understood concept of transient Markov chains. We enumerate crucial characteristics of this objective and define the notion of universally transient MDPs.

Fix a countable MDP $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P)$. Define the transience objective, denoted by `Transience`, to be the set of runs that do not visit any state of $\mathcal{M}$ infinitely often, i.e.,

$$\texttt{Transience} \stackrel{\text{def}}{=} \bigwedge_{s \in S} \mathsf{FG}\, \neg s.$$

The `Transience` objective is tail, as it is closed under removing finite prefixes of runs. Also note that `Transience` cannot be encoded in a parity objective.

We call $\mathcal{M}$ *universally transient* iff for all states s_0, for all strategies σ, the `Transience` property holds almost-surely from s_0, i.e.,

$$\forall s_0 \in S \;\; \forall \sigma \in \Sigma \;\; \mathcal{P}_{\mathcal{M},s_0,\sigma}(\texttt{Transience}) = 1.$$

The MDP in Figure 1 models the classical Gambler's Ruin Problem with restart; see [10, Chapter 14]. It is well-known that if the controller starts with wealth i and if $p \leq \frac{1}{2}$, the probability of ruin (visiting the state w_0) is $\mathcal{P}_{w_i}(\mathsf{F}\, w_0) = 1$. Consequently, the probability of re-visiting w_0 infinitely often is 1, implying that $\mathcal{P}_{w_i}(\texttt{Transience}) = 0$. In contrast, for the case with $p > \frac{1}{2}$, for all states w_i, the probability of re-visiting w_i is strictly below 1. Hence, the `Transience` property holds almost-surely. This example indicates that the transience property depends on the probability values of the transitions and not just on the underlying transition graph, and thus may require arithmetic reasoning. In particular, the MDP in Figure 1 is universally transient iff $p > \frac{1}{2}$.

In general, optimal strategies for `Transience` need not exist:

▶ **Lemma 1.** *There exists a finitely branching countable MDP with initial state s_0 such that*
- $\texttt{val}_{\texttt{Transience}}(s) = 1$ *for all controlled states s,*
- *there does not exist any optimal strategy σ such that $\mathcal{P}_{s_0,\sigma}(\texttt{Transience}) = 1$.*

Proof. Consider a countable MDP $\mathcal{M}$ with set $S = \{\ell_i, \ell'_i, r_i, x_i \mid i \geq 1\} \cup \{\ell_0, \bot\}$ of states; see Figure 2. For all $i \geq 1$ the state x_{i+1} is the unique successor of x_i so that $(x_i)_{i\geq 1}$ form an acyclic ladder; the value of `Transience` is 1 for all x_i. The state $\bot$ is sink, and its value is 0. The states $(r_i)_{i\geq 1}$ are all random, and $r_i \xrightarrow{1-2^{-i}} x_i$ and $r_i \xrightarrow{2^{-i}} \bot$. Observe that the value of `Transience` is $1 - 2^{-i}$ for the r_i.

The states $(\ell_i)_{i\in\mathbb{N}}$ are controlled whereas the states $(\ell'_i)_{i\geq 1}$ are random. By interleaving of these states, we construct a "recurrent ladder" of decisions: $\ell_0 \to \ell_1$ and for all $i \geq 1$, state ℓ_i has two successors ℓ'_i and r_i. In random states ℓ'_i, as in Gambler's Ruin with a fair coin, the successors are ℓ_{i-1} or ℓ_{i+1}, each with equal probability. In each state $(\ell_i)_{i\geq 1}$, the controller decides to either stay on the ladder by going to ℓ'_i or leaves the ladder to r_i. As in Figure 1, if the controller stays on the ladder forever, the probability of `Transience` is 0.

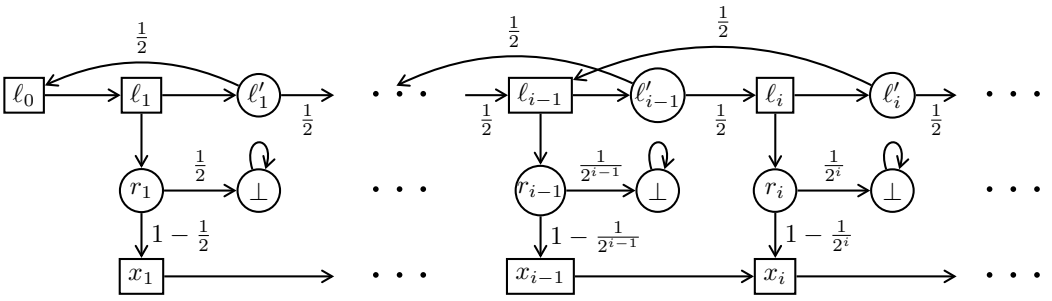

Figure 2 A partial illustration of the MDP in Lemma 1, in which there is no optimal strategy for `Transience`, starting from states ℓ_i. For readability, we have three copies of the state $\perp$. We call the ladder consisting of the interleaved controlled states ℓ_i and random states ℓ'_i a "recurrent ladder": if the controller stays on this ladder forever, it faithfully simulates a Gambler's Ruin with a fair coin, and the probability of `Transience` will be 0.

Starting in ℓ_0, for all $i > 0$, strategy σ_i that stays on the ladder until visiting ℓ_i (which happens eventually almost surely) and then leaves the ladder to r_i achieves `Transience` with probability $1 - 2^i$. Hence, $\mathtt{val}_{\mathtt{Transience}}(\ell_0) = 1$.

Recall that transience cannot be achieved with a positive probability by staying on the acyclic ladder forever. But any strategy that leaves the ladder with a positive probability comes with a positive probability of falling into $\perp$, thus is not optimal either. Thus there is no optimal strategy for `Transience`. ◀

Reduction to Finitely Branching MDPs. In our main results, we will prove that for the `Transience` property there always exist ε-optimal MD strategies in finitely branching countable MDPs; and if an optimal strategy exists, there will exist an optimal MD strategy. We generalize these results to infinitely branching countable MDPs by the following reduction:

▶ **Lemma 2.** *Given an infinitely branching countable MDP $\mathcal{M}$ with an initial state s_0, there exists a finitely branching countable $\mathcal{M}'$ with a set S' of states such that $s_0 \in S'$ and*

1. *each strategy α_1 in $\mathcal{M}$ is mapped to a unique strategy β_1 in $\mathcal{M}'$ where*

$$\mathcal{P}_{s_0,\alpha_1}(\mathtt{Transience}) = \mathcal{P}_{s_0,\beta_1}(\mathtt{Transience}),$$

2. *and conversely, every MD strategy β_2 in $\mathcal{M}'$ is mapped to an MD strategy α_2 in $\mathcal{M}$ where*

$$\mathcal{P}_{s_0,\alpha_2}(\mathtt{Transience}) \geq \mathcal{P}_{s_0,\beta_2}(\mathtt{Transience}).$$

Properties of Universally Transient MDPs. Notice that acyclicity implies universal transience, but not vice-versa.

▶ **Lemma 3.** *For every countable MDP $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P)$, the following conditions are equivalent.*

1. *$\mathcal{M}$ is universally transient, i.e., $\forall s_0, \forall \sigma. \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathtt{Transience}) = 1$.*
2. *For every initial state s_0 and state s, the objective of re-visiting s infinitely often has value zero, i.e., $\forall s_0, s \sup_\sigma \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathsf{GF}(s)) = 0$.*
3. *For every state s the value of the objective to re-visit s is strictly below 1, i.e., $Re(s) \stackrel{def}{=} \sup_\sigma \mathcal{P}_{\mathcal{M},s,\sigma}(\mathsf{XF}(s)) < 1$.*

4. *For every state s there exists a finite bound $B(s)$ such that for every state s_0 and strategy σ from s_0 the expected number of visits to s is $\leq B(s)$.*
5. *For all states s_0, s, under every strategy σ from s_0 the expected number of visits to s is finite.*

Proof. Towards (1) $\Rightarrow$ (2), consider an arbitrary strategy σ from the initial state s_0 and some state s. By (1) we have $\forall\sigma.\mathcal{P}_{\mathcal{M},s_0,\sigma}(\texttt{Transience}) = 1$ and thus $0 = \mathcal{P}_{\mathcal{M},s_0,\sigma}(\neg\texttt{Transience}) = \mathcal{P}_{\mathcal{M},s_0,\sigma}(\bigcup_{s'\in S}\mathsf{GF}(s')) \geq \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathsf{GF}(s))$ which implies (2).

Towards (2) $\Rightarrow$ (1), consider an arbitrary strategy σ from the initial state s_0. By (2) we have $0 = \sum_{s\in S}\mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathsf{GF}(s)) \geq \mathcal{P}_{\mathcal{M},s_0,\sigma}(\bigcup_{s\in S}\mathsf{GF}(s)) = \mathcal{P}_{\mathcal{M},s_0,\sigma}(\neg\texttt{Transience})$ and thus $\mathcal{P}_{\mathcal{M},s_0,\sigma}(\texttt{Transience}) = 1$.

We now show the implications (2) $\Rightarrow$ (3) $\Rightarrow$ (4) $\Rightarrow$ (5) $\Rightarrow$ (2).

Towards $\neg(3) \Rightarrow \neg(2)$, $\neg(3)$ implies $\exists s.Re(s) = 1$ and thus $\forall\varepsilon > 0.\exists\sigma_\varepsilon\,\mathcal{P}_{\mathcal{M},s,\sigma_\varepsilon}(\mathsf{XF}(s)) \geq 1-\varepsilon$. Let $\varepsilon_i \stackrel{\text{def}}{=} 2^{-(i+1)}$. We define the strategy σ to play like σ_{ε_i} between the i-th and $(i+1)$th visit to s. Since $\sum_{i=1}^{\infty}\varepsilon_i < \infty$, we have $\prod_{i=1}^{\infty}(1-\varepsilon_i) > 0$. Therefore $\mathcal{P}_{\mathcal{M},s,\sigma}(\mathsf{GF}(s)) \geq \prod_{i=1}^{\infty}(1-\varepsilon_i) > 0$, which implies $\neg(2)$, where $s_0 = s$.

Towards (3) $\Rightarrow$ (4), regardless of s_0 and the chosen strategy, the expected number of visits to s is upper-bounded by $B(s) \stackrel{\text{def}}{=} \sum_{n=0}^{\infty}(n+1)\cdot(Re(s))^n < \infty$.

The implication (4) $\Rightarrow$ (5) holds trivially.

Towards $\neg(2) \Rightarrow \neg(5)$, by $\neg(2)$ there exist states s_0, s and a strategy σ such that $\mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathsf{GF}(s)) > 0$. Thus the expected number of visits to s is infinite, which implies $\neg(5)$. ◀

We remark that if an MDP is *not* universally transient (unlike in Lemma 3(5)), for a strategy σ, the expected number of visits to some state can be infinite, even if σ attains `Transience` almost surely.

Consider the MDP $\mathcal{M}$ with controlled states $\{s_0, s_1, \dots\}$, initial state s_0 and transitions $s_0 \to s_0$ and $s_k \to s_{k+1}$ for every $k \geq 0$. We define a strategy σ that, while in state s_0, proceeds in rounds $i = 1, 2, \dots$. In the i-th round it tosses a fair coin. If Heads then it goes to s_1. If Tails then it loops around s_0 exactly 2^i times and then goes to round $i+1$. In every round the probability of going to s_1 is $1/2$ and therefore the probability of staying in s_0 forever is $(1/2)^\infty = 0$. Thus $\mathcal{P}_{\mathcal{M},s_0,\sigma}(\texttt{Transience}) = 1$. However, the expected number of visits to s_0 is $\geq \sum_{i=1}^{\infty}\left(\frac{1}{2}\right)^i \cdot 2^i = \infty$.

4 MD Strategies for Transience

We show that there exist uniformly ε-optimal MD strategies for `Transience` and that optimal strategies, where they exist, can also be chosen MD.

First we show that there exist ε-optimal deterministic 1-bit strategies for `Transience` (in Corollary 5) and then we show how to dispense with the 1-bit memory (in Lemma 6).

It was shown in [14] that there exist ε-optimal deterministic 1-bit strategies for Büchi objectives in *acyclic* countable MDPs (though not in general MDPs). These 1-bit strategies will be similar to the 1-bit strategies for `Transience` that we aim for in (not necessarily acyclic) countable MDPs. In Lemma 4 below we first strengthen the result from [14] and construct ε-optimal deterministic 1-bit strategies for objectives $\texttt{Büchi}(F) \cap \texttt{Transience}$. From this we obtain deterministic 1-bit strategies for `Transience` (Corollary 5).

▶ **Lemma 4.** *Let $\mathcal{M}$ be a countable MDP, I a finite set of initial states, F a set of states and $\varepsilon > 0$. Then there exists a deterministic 1-bit strategy for $\texttt{Büchi}(F) \cap \texttt{Transience}$ that is ε-optimal from every $s \in I$.*

Proof sketch. It follows the proof of [14, Theorem 5], which considers $\texttt{Büchi}(F)$ conditions for *acyclic* (and hence universally transient) MDPs. The only part of that proof that requires modification is [14, Lemma 10], which is replaced here by [13, Lemma 18] to deal with general MDPs.

In short, from every $s \in I$ there exists an ε-optimal strategy σ_s for $\varphi \stackrel{\text{def}}{=} \texttt{Büchi}(F) \cap \texttt{Transience}$. We 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 [13, Lemma 18], we can define a second stronger objective $\varphi' \subseteq \varphi$ and show $\forall_{s \in I} \mathcal{P}_{\mathcal{M},s,\sigma_s}(\varphi') \geq \texttt{val}_{\mathcal{M},\varphi}(s) - 2\varepsilon$. We then construct a deterministic 1-bit strategy σ' that is optimal for φ' from all $s \in I$ and thus 2ε-optimal for φ. Since ε can be chosen arbitrarily small, the result follows. ◀

Unlike for the $\texttt{Transience}$ objective alone (see below), the 1-bit memory is strictly necessary for the $\texttt{Büchi}(F) \cap \texttt{Transience}$ objective in Lemma 4. The 1-bit lower bound for $\texttt{Büchi}(F)$ objectives in [14] holds even for acyclic MDPs where $\texttt{Transience}$ is trivially true.

▶ **Corollary 5.** *Let $\mathcal{M}$ be a countable MDP, I a finite set of initial states, F a set of states and $\varepsilon > 0$.*

1. *If $\forall s \in I\ \texttt{val}_{\mathcal{M},\texttt{Büchi}(F)}(s) = \texttt{val}_{\mathcal{M},\texttt{Büchi}(F)\cap\texttt{Transience}}(s)$ then there exists a deterministic 1-bit strategy for $\texttt{Büchi}(F)$ that is ε-optimal from every $s \in I$.*
2. *If $\mathcal{M}$ is universally transient then there exists a deterministic 1-bit strategy for $\texttt{Büchi}(F)$ that is ε-optimal from every $s \in I$.*
3. *There exists a deterministic 1-bit strategy for $\texttt{Transience}$ that is ε-optimal from every $s \in I$.*

Proof. Towards (1), since $\forall s \in I\ \texttt{val}_{\mathcal{M},\texttt{Büchi}(F)}(s) = \texttt{val}_{\mathcal{M},\texttt{Büchi}(F)\cap\texttt{Transience}}(s)$, strategies that are ε-optimal for $\texttt{Büchi}(F) \cap \texttt{Transience}$ are also ε-optimal for $\texttt{Büchi}(F)$. Thus the result follows from Lemma 4.

Item (2) follows directly from (1), since the precondition always holds in universally transient MDPs.

Towards (3), let $F \stackrel{\text{def}}{=} S$. Then we have $\texttt{Büchi}(F) \cap \texttt{Transience} = \texttt{Transience}$ and we obtain from Lemma 4 that there exists a deterministic 1-bit strategy for $\texttt{Transience}$ that is ε-optimal from every $s \in I$. ◀

Note that every acyclic MDP is universally transient and thus Corollary 5(2) implies the upper bound on the strategy complexity of $\texttt{Büchi}(F)$ from [14] (but not vice-versa).

In the next step we show how to dispense with the 1-bit memory and obtain non-uniform ε-optimal MD strategies for $\texttt{Transience}$.

▶ **Lemma 6.** *Let $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P)$ be a countable MDP with initial state s_0, and $\varepsilon > 0$. There exists an MD strategy σ that is ε-optimal for $\texttt{Transience}$ from s_0, i.e., $\mathcal{P}_{\mathcal{M},s_0,\sigma}(\texttt{Transience}) \geq \texttt{val}_{\mathcal{M},\texttt{Transience}}(s_0) - \varepsilon$.*

Proof. By Lemma 2 it suffices to prove the property for finitely branching MDPs. Thus without restriction in the rest of the proof we assume that $\mathcal{M}$ is finitely branching.

Let $\varepsilon' \stackrel{\text{def}}{=} \varepsilon/2$. We instantiate Corollary 5(3) with $I \stackrel{\text{def}}{=} \{s_0\}$ and obtain that there exists an ε'-optimal deterministic 1-bit strategy $\hat{\sigma}$ for $\texttt{Transience}$ from s_0.

We now construct a slightly modified MDP $\mathcal{M}'$ as follows. Let $S_{bad} \subseteq S$ be the subset of states where $\hat{\sigma}$ attains zero for $\texttt{Transience}$ in *both* memory modes, i.e., $S_{bad} \stackrel{\text{def}}{=} \{s \in S \mid \mathcal{P}_{\mathcal{M},s,\sigma[0]}(\texttt{Transience}) = \mathcal{P}_{\mathcal{M},s,\sigma[1]}(\texttt{Transience}) = 0\}$. Let $S_{good} \stackrel{\text{def}}{=} S \setminus S_{bad}$. We obtain $\mathcal{M}'$ from $\mathcal{M}$ by making all states in S_{bad} losing sinks (for $\texttt{Transience}$), by deleting all outgoing edges and adding a self-loop instead. It follows that

$$\mathcal{P}_{\mathcal{M},s_0,\hat{\sigma}}(\texttt{Transience}) = \mathcal{P}_{\mathcal{M}',s_0,\hat{\sigma}}(\texttt{Transience}) \tag{1}$$

$$\forall \sigma.\ \mathcal{P}_{\mathcal{M},s_0,\sigma}(\texttt{Transience}) \geq \mathcal{P}_{\mathcal{M}',s_0,\sigma}(\texttt{Transience}) \tag{2}$$

In the following we show that it is possible to play in such a way that, for every $s \in S_{good}$, the expected number of visits to s is *finite*. We obtain the deterministic 1-bit strategy σ' in $\mathcal{M}'$ by modifying $\hat{\sigma}$ as follows. In every state s and memory mode $x \in \{0,1\}$ where $\hat{\sigma}[x]$ attains 0 for $\texttt{Transience}$ and $\hat{\sigma}[1-x]$ attains > 0 the strategy σ' sets the memory bit to $1-x$. (Note that only states $s \in S_{good}$ can be affected by this change.) It follows that

$$\forall s \in S.\ \mathcal{P}_{\mathcal{M}',s,\sigma'}(\texttt{Transience}) \geq \mathcal{P}_{\mathcal{M}',s,\hat{\sigma}}(\texttt{Transience}) \tag{3}$$

Moreover, from all states in S_{good} in $\mathcal{M}'$ the strategy σ' attains a strictly positive probability of $\texttt{Transience}$ in *both* memory modes, i.e., for all $s \in S_{good}$ we have

$$t(s,\sigma') \stackrel{\text{def}}{=} \min_{x\in\{0,1\}} \mathcal{P}_{\mathcal{M}',s,\sigma'[i]}(\texttt{Transience}) > 0.$$

Let $r(s,\sigma',x)$ be the probability, when playing $\sigma'[x]$ from state s, of reaching s again in the *same* memory mode x. For every $s \in S_{good}$ we have $r(s,\sigma',x) < 1$, since $t(s,\sigma') > 0$.

Let $R(s)$ be the expected number of visits to state s when playing σ' from s_0 in $\mathcal{M}'$, and $R_x(s)$ the expected number of visits to s in memory mode $x \in \{0,1\}$. For all $s \in S_{good}$ we have that

$$R(s) = R_0(s) + R_1(s) \leq \sum_{n=1}^{\infty} n \cdot r(s,\sigma',0)^{n-1} + \sum_{n=1}^{\infty} n \cdot r(s,\sigma',1)^{n-1} < \infty \tag{4}$$

where the first equality holds by linearity of expectations. Thus the expected number of visits to s is *finite*.

Now we upper-bound the probability of visiting S_{bad}. We have $\mathcal{P}_{\mathcal{M}',s_0,\sigma'}(\texttt{Transience}) \geq \mathcal{P}_{\mathcal{M}',s_0,\hat{\sigma}}(\texttt{Transience}) = \mathcal{P}_{\mathcal{M},s_0,\hat{\sigma}}(\texttt{Transience}) \geq \texttt{val}_{\mathcal{M},\texttt{Transience}}(s_0) - \varepsilon'$ by (3), (1) and the ε'-optimality of $\hat{\sigma}$. Since states in S_{bad} are losing sinks in $\mathcal{M}'$, it follows that

$$\mathcal{P}_{\mathcal{M}',s_0,\sigma'}(\mathsf{F} S_{bad}) \leq 1 - \mathcal{P}_{\mathcal{M}',s_0,\sigma'}(\texttt{Transience}) \leq 1 - \texttt{val}_{\mathcal{M},\texttt{Transience}}(s_0) + \varepsilon' \tag{5}$$

We now augment the MDP $\mathcal{M}'$ by assigning costs to transitions as follows. Let $i : S \to \mathbb{N}$ be an enumeration of the state space, i.e., a bijection. Let $S'_{good} \stackrel{\text{def}}{=} \{s \in S_{good} \mid R(s) > 0\}$ be the subset of states in S_{good} that are visited with non-zero probability when playing σ' from s_0. Each transition $s' \to s$ is assigned a cost:

- If $s' \in S_{bad}$ then $s \in S_{bad}$ by def. of $\mathcal{M}'$. We assign cost 0.
- If $s' \in S_{good}$ and $s \in S_{bad}$ we assign cost $K/(1 - \texttt{val}_{\mathcal{M},\texttt{Transience}}(s_0) + \varepsilon')$ for $K \stackrel{\text{def}}{=} (1+\varepsilon')/\varepsilon'$.
- If $s' \in S_{good}$ and $s \in S'_{good}$ we assign cost $2^{-i(s)}/R(s)$. This is well defined, since $R(s) > 0$.
- $s' \in S_{good}$ and $s \in S_{good} \setminus S'_{good}$ we assign cost 1.

Note that all transitions leading to states in S_{good} are assigned a non-zero cost, since $R(s)$ is finite by (4).

When playing σ' from s_0 in $\mathcal{M}'$, the expected total cost is upper-bounded by

$$\mathcal{P}_{\mathcal{M}',s_0,\sigma'}(\mathsf{F} S_{bad}) \cdot K/(1 - \texttt{val}_{\mathcal{M},\texttt{Transience}}(s_0) + \varepsilon') + \sum_{s\in S'_{good}} R(s) \cdot 2^{-i(s)}/R(s)$$

The first part is $\leq K$ by (5) and the second part is ≤ 1, since $R(s) < \infty$ by (4). Therefore the expected total cost is $\leq K+1$, i.e., σ' witnesses that it is possible to attain a finite expected cost that is upper-bounded by $K+1$.

Now we define our MD strategy σ. Let σ be an optimal MD strategy on $\mathcal{M}'$ (from s_0) that minimizes the expected cost. It exists, as a finite expected cost is attainable and $\mathcal{M}'$ is finitely branching; see [21, Theorem 7.3.6].

We now show that σ attains $\texttt{Transience}$ with high probability in $\mathcal{M}'$ (and in $\mathcal{M}$). Since σ is cost-optimal, its attained cost from s_0 is upper-bounded by that of σ', i.e., $\leq K+1$. Since the cost of entering S_{bad} is $K/(1-\texttt{val}_{\mathcal{M},\texttt{Transience}}(s_0)+\varepsilon')$, we have $\mathcal{P}_{\mathcal{M}',s_0,\sigma}(\mathsf{F}S_{bad}) \cdot K/(1-\texttt{val}_{\mathcal{M},\texttt{Transience}}(s_0)+\varepsilon') \leq K+1$ and thus

$$\mathcal{P}_{\mathcal{M}',s_0,\sigma}(\mathsf{F}S_{bad}) \leq \frac{K+1}{K}(1-\texttt{val}_{\mathcal{M},\texttt{Transience}}(s_0)+\varepsilon') \tag{6}$$

For every state $s \in S_{good}$, all transitions into s have the same fixed non-zero cost. Thus every run that visits some state $s \in S_{good}$ infinitely often has infinite cost. Since the expected cost of playing σ from s_0 is $\leq K+1$, such runs must be a null-set, i.e.,

$$\mathcal{P}_{\mathcal{M}',s_0,\sigma}(\neg\texttt{Transience} \wedge \mathsf{G}S_{good}) = 0 \tag{7}$$

Thus

$$\begin{aligned}
&\mathcal{P}_{\mathcal{M},s_0,\sigma}(\texttt{Transience}) \\
&\geq \mathcal{P}_{\mathcal{M}',s_0,\sigma}(\texttt{Transience}) && \text{by (2)} \\
&= 1-\mathcal{P}_{\mathcal{M}',s_0,\sigma}(\mathsf{F}S_{bad}) && \text{by (7)} \\
&\geq 1-\frac{K+1}{K}(1-\texttt{val}_{\mathcal{M},\texttt{Transience}}(s_0)+\varepsilon') && \text{by (6)} \\
&= \texttt{val}_{\mathcal{M},\texttt{Transience}}(s_0)-\varepsilon'-(1/K)(1-\texttt{val}_{\mathcal{M},\texttt{Transience}}(s_0)+\varepsilon') \\
&\geq \texttt{val}_{\mathcal{M},\texttt{Transience}}(s_0)-\varepsilon'-(1/K)(1+\varepsilon') \\
&= \texttt{val}_{\mathcal{M},\texttt{Transience}}(s_0)-2\varepsilon' && \text{def. of } K \\
&= \texttt{val}_{\mathcal{M},\texttt{Transience}}(s_0)-\varepsilon && \text{def. of } \varepsilon' \qquad \blacktriangleleft
\end{aligned}$$

Now we lift the result of Lemma 6 from non-uniform to uniform strategies (and to optimal strategies) and obtain the following theorem. The proof is a generalization of a "plastering" construction by Ornstein [20] (see also [16]) from reachability to tail objectives, which works by fixing MD strategies on ever expanding subsets of the state space.

▶ **Theorem 7.** *Let $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P)$ be a countable MDP, and let φ be an objective that is tail in $\mathcal{M}$. Suppose for every $s \in S$ there exist ε-optimal MD strategies for φ. Then:*

1. *There exist uniform ε-optimal MD strategies for φ.*
2. *There exists a single MD strategy that is optimal from every state that has an optimal strategy.*

▶ **Theorem 8.** *In every countable MDP there exist uniform ε-optimal MD strategies for* $\texttt{Transience}$*. Moreover, there exists a single MD strategy that is optimal for* $\texttt{Transience}$ *from every state that has an optimal strategy.*

Proof. Immediate from Lemma 6 and Theorem 7, since $\texttt{Transience}$ is a tail objective. ◀

5 Strategy Complexity in Universally Transient MDPs

The strategy complexity of parity objectives in general MDPs is known [15]. Here we show that some parity objectives have a lower strategy complexity in universally transient MDPs. It is known [14] that there are acyclic (and hence universally transient) MDPs where ε-optimal strategies for $\{1,2\}$-$\mathtt{Parity}$ (and optimal strategies for $\{1,2,3\}$-$\mathtt{Parity}$, resp.) require 1 bit.

We show that, for all simpler parity objectives in the Mostowski hierarchy [19], universally transient MDPs admit uniformly (ε-)optimal MD strategies (unlike general MDPs [15]). These results (Theorems 10 and 11) ultimately rely on the existence of uniformly ε-optimal strategies for safety objectives. While such strategies always exist for finitely branching MDPs – simply pick a value-maximal successor – this is not the case for infinitely branching MDPs [17]. However, we show that universal transience implies the existence of uniformly ε-optimal strategies for safety objectives even for *infinitely branching* MDPs.

▶ **Theorem 9.** *For every universally transient countable MDP, safety objective and $\varepsilon > 0$ there exists a uniformly ϵ-optimal MD strategy.*

Proof. Let $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P)$ be a universally transient MDP and $\varepsilon > 0$. Assume w.l.o.g. that the target $T \subseteq S$ of the objective $\varphi = \mathtt{Safety}(T)$ is a (losing) sink and let $\iota : S \to \mathbb{N}$ be an enumeration of the state space S.

By Lemma 3(3), for every state s we have $Re(s) \stackrel{\text{def}}{=} \sup_\sigma \mathcal{P}_{\mathcal{M},s,\sigma}(\mathsf{XF}(s)) < 1$ and thus $R(s) \stackrel{\text{def}}{=} \sum_{i=0}^{\infty} Re(s)^i < \infty$. This means that, independent of the chosen strategy, $Re(s)$ upper-bounds the chance to return to s, and $R(s)$ bounds the expected number of visits to s.

Suppose that σ is an MD strategy which, at any state $s \in S_\Box$, picks a successor s' with

$$\mathtt{val}(s') \quad \geq \quad \mathtt{val}(s) - \frac{\varepsilon}{2^{\iota(s)+1} \cdot R(s)}.$$

This is possible even if $\mathcal{M}$ is infinitely branching, by the definition of value and the fact that $R(s) < \infty$. We show that $\mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathtt{Safety}(T)) \geq \mathtt{val}(s_0) - \varepsilon$ holds for every initial state s_0, which implies the claim of the theorem.

Towards this, we define a function $\mathtt{cost}$ that labels each transition in the MDP with a real-valued cost: For every controlled transition $s \longrightarrow s'$ let $\mathtt{cost}((s,s')) \stackrel{\text{def}}{=} \mathtt{val}(s) - \mathtt{val}(s') \geq 0$. Random transitions have cost zero. We will argue that when playing σ from any start state s_0, its attainment w.r.t. the objective $\mathtt{Safety}(T)$ equals the value of s_0 minus the expected total cost, and that this cost is bounded by ε.

For any $i \in \mathbb{N}$ let us write s_i for the random variable denoting the state just after step i, and $\mathtt{Cost}(i) \stackrel{\text{def}}{=} \mathtt{cost}(s_i, s_{i+1})$ for the cost of step i in a random run. We observe that under σ the expected total cost is bounded in the limit, i.e.,

$$\lim_{n\to\infty} \mathcal{E}\left(\sum_{i=0}^{n-1} \mathtt{Cost}(i)\right) \leq \varepsilon. \tag{8}$$

We moreover note that for every n,

$$\mathcal{E}(\mathtt{val}(s_n)) = \mathcal{E}(\mathtt{val}(s_0)) - \mathcal{E}\left(\sum_{i=0}^{n-1} \mathtt{Cost}(i)\right). \tag{9}$$

Full proofs of the above two equations can be found in [13]. Together they imply

$$\liminf_{n\to\infty} \mathcal{E}(\mathtt{val}(s_n)) = \mathtt{val}(s_0) - \lim_{n\to\infty} \mathcal{E}\left(\sum_{i=0}^{n-1} \mathtt{cost}(i)\right) \geq \mathtt{val}(s_0) - \varepsilon. \tag{10}$$

Finally, to show the claim let $[s_n \notin T] : S^\omega \to \{0,1\}$ be the random variable that indicates that the n-th state is not in the target set T. Note that $[s_n \notin T] \geq \mathtt{val}(s_n)$ because target states have value 0. We have:

$$\begin{aligned}
\mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathtt{Safety}(T)) &= \mathcal{P}_{\mathcal{M},s_0,\sigma}\left(\bigwedge_{i=0}^{\infty} \mathsf{X}^i \neg T\right) && \text{semantics of } \mathtt{Safety}(T) = \mathsf{G}\neg T \\
&= \lim_{n\to\infty} \mathcal{P}_{\mathcal{M},s_0,\sigma}\left(\bigwedge_{i=0}^{n} \mathsf{X}^i \neg T\right) && \text{continuity of measures} \\
&= \lim_{n\to\infty} \mathcal{P}_{\mathcal{M},s_0,\sigma}(\mathsf{X}^n \neg T) && T \text{ is a sink} \\
&= \lim_{n\to\infty} \mathcal{E}([s_n \notin T]) && \text{definition of } [s_n \notin T] \\
&\geq \liminf_{n\to\infty} \mathcal{E}(\mathtt{val}(s_n)) && \text{as } [s_n \notin T] \geq \mathtt{val}(s_n) \\
&\geq \mathtt{val}(s_0) - \varepsilon && \text{Equation (10).}
\end{aligned}$$

◀

We can now combine Theorem 9 with the results from [15] to show the existence of MD strategies assuming universal transience.

▶ **Theorem 10.** *For universally transient MDPs optimal strategies for* $\{0,1,2\}$-$\mathtt{Parity}$*, where they exist, can be chosen uniformly MD.*

Formally, let $\mathcal{M}$ *be a universally transient MDP with states* S*,* $Col : S \to \{0,1,2\}$*, and* $\varphi = \mathtt{Parity}(Col)$*. There exists an MD strategy* σ' *that is optimal for all states* s *that have an optimal strategy:* $\left(\exists \sigma \in \Sigma. \mathcal{P}_{\mathcal{M},s,\sigma}(\varphi) = \mathtt{val}_{\mathcal{M}}(s)\right) \implies \mathcal{P}_{\mathcal{M},s,\sigma'}(\varphi) = \mathtt{val}_{\mathcal{M}}(s)$*.*

Proof. Let $\mathcal{M}_+$ be the conditioned version of $\mathcal{M}$ w.r.t. φ (see [15, Def. 19] for a precise definition). By Lemma 17, $\mathcal{M}_+$ is still a universally transient MDP and therefore by Theorem 9, there exist uniformly ε-optimal MD strategies for every safety objective and every $\varepsilon > 0$. The claim now follows from [15, Theorem 22]. ◀

▶ **Theorem 11.** *For every universally transient countable MDP* $\mathcal{M}$*, co-Büchi objective and* $\varepsilon > 0$ *there exists a uniformly* ε*-optimal MD strategy.*

Formally, let $\mathcal{M}$ *be a universally transient countable MDP with states* S*,* $Col : S \to \{0,1\}$ *be a coloring,* $\varphi = \mathtt{Parity}(Col)$ *and* $\varepsilon > 0$*.*

There exists an MD strategy σ' *s.t. for every state* s*,* $\mathcal{P}_{\mathcal{M},s,\sigma'}(\varphi) \geq \mathtt{val}_{\mathcal{M}}(s) - \varepsilon$*.*

Proof. This directly follows from Theorem 9 and [15, Theorem 25]. ◀

6 The Conditioned MDP

Given an MDP $\mathcal{M}$ and an objective φ that is tail in $\mathcal{M}$, a construction of a *conditioned* MDP $\mathcal{M}_+$ was provided in [17, Lemma 6] that, very loosely speaking, "scales up" the probability of φ so that any strategy σ is optimal in $\mathcal{M}$ if it is almost surely winning in $\mathcal{M}_+$. For certain tail objectives, this construction was used in [17] to reduce the sufficiency of MD strategies for *optimal* strategies to the sufficiency of MD strategies for *almost surely winning* strategies, which is a special case that may be easier to handle.

However, the construction was restricted to states that *have* an optimal strategy. In fact, states in $\mathcal{M}$ that do not have an optimal strategy do not appear in $\mathcal{M}_+$. In the following, we lift this restriction by constructing a more general version of the conditioned MDP, called $\mathcal{M}_*$. The MDP $\mathcal{M}_*$ will contain all states from $\mathcal{M}$ that have a positive value w.r.t. φ in $\mathcal{M}$. Moreover, all these states will have value 1 in $\mathcal{M}_*$. It will then follow from Lemma 13(3) below that an ε-optimal strategy in $\mathcal{M}_*$ is $\varepsilon\mathtt{val}_{\mathcal{M}}(s_0)$-optimal in $\mathcal{M}$. This allows us to reduce the sufficiency of MD strategies for ε-optimal strategies to the sufficiency of MD

strategies for ε-optimal strategies for states with value 1. In fact, it also follows that if an MD strategy σ is uniform ε-optimal in $\mathcal{M}_*$, it is *multiplicatively* uniform ε-optimal in $\mathcal{M}$, i.e., $\mathcal{P}_{\mathcal{M},s,\sigma}(\varphi) \geq (1-\varepsilon) \cdot \mathtt{val}_{\mathcal{M}}(s)$ holds for all states s.

▶ **Definition 12.** *For an MDP $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P)$ and an objective φ that is tail in $\mathcal{M}$, 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*

$$
\begin{aligned}
S_{*\Box} &= \{s \in S_\Box \mid \mathtt{val}_{\mathcal{M}}(s) > 0\} \\
S_{*\bigcirc} &= \{s \in S_\bigcirc \mid \mathtt{val}_{\mathcal{M}}(s) > 0\} \cup \{s_\bot\} \cup \{(s,t) \in \longrightarrow \mid s \in S_\Box,\ \mathtt{val}_{\mathcal{M}}(s) > 0\} \\
\longrightarrow_* &= \{(s,(s,t)) \in (S_\Box \times \longrightarrow) \mid \mathtt{val}_{\mathcal{M}}(s) > 0,\ s \longrightarrow t\} \cup \\
&\quad \{(s,t) \in S_\bigcirc \times S \mid \mathtt{val}_{\mathcal{M}}(s) > 0,\ \mathtt{val}_{\mathcal{M}}(t) > 0\} \cup \\
&\quad \{((s,t),t) \in (\longrightarrow \times S) \mid \mathtt{val}_{\mathcal{M}}(s) > 0,\ \mathtt{val}_{\mathcal{M}}(t) > 0\} \cup \\
&\quad \{((s,t),s_\bot) \in (\longrightarrow \times \{s_\bot\}) \mid \mathtt{val}_{\mathcal{M}}(s) > \mathtt{val}_{\mathcal{M}}(t)\} \cup \\
&\quad \{(s_\bot, s_\bot)\}
\end{aligned}
$$

$$
P_*(s,t) = P(s,t) \cdot \frac{\mathtt{val}_{\mathcal{M}}(t)}{\mathtt{val}_{\mathcal{M}}(s)} \qquad P_*((s,t),t) = \frac{\mathtt{val}_{\mathcal{M}}(t)}{\mathtt{val}_{\mathcal{M}}(s)}
$$

$$
P_*((s,t),s_\bot) = 1 - \frac{\mathtt{val}_{\mathcal{M}}(t)}{\mathtt{val}_{\mathcal{M}}(s)} \qquad P_*(s_\bot, s_\bot) = 1
$$

for a fresh state $s_\bot$.

The conditioned MDP is well-defined. Indeed, as φ is tail in $\mathcal{M}$, for any $s \in S_\bigcirc$ we have $\mathtt{val}_{\mathcal{M}}(s) = \sum_{s \longrightarrow t} P(s,t)\mathtt{val}_{\mathcal{M}}(t)$, and so if $\mathtt{val}_{\mathcal{M}}(s) > 0$ then $\sum_{s \longrightarrow t} P_*(s,t) = 1$.

▶ **Lemma 13.** *Let $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P)$ be an MDP, and let φ be an objective that is tail in $\mathcal{M}$. Let $\mathcal{M}_* = (S_*, S_{*\Box}, S_{*\bigcirc}, \longrightarrow_*, P_*)$ be the conditioned version of $\mathcal{M}$ w.r.t. φ. Let $s_0 \in S_* \cap S$. Let $\sigma \in \Sigma_{\mathcal{M}_*}$, and note that σ can be transformed to a strategy in $\mathcal{M}$ in a natural way. Then:*

1. *For all $n \geq 0$ and all partial runs $s_0 s_1 \cdots s_n \in s_0 S_*^*$ in $\mathcal{M}_*$ with $s_n \in S$:*

$$
\mathtt{val}_{\mathcal{M}}(s_0) \cdot \mathcal{P}_{\mathcal{M}_*,s_0,\sigma}(s_0 s_1 \cdots s_n S_*^\omega) = \mathcal{P}_{\mathcal{M},s_0,\sigma}(\overline{s_0 s_1 \cdots s_n} S^\omega) \cdot \mathtt{val}_{\mathcal{M}}(s_n)\,,
$$

   where $\overline{w}$ for a partial run w in $\mathcal{M}_*$ refers to its natural contraction to a partial run in $\mathcal{M}$; i.e., $\overline{w}$ is obtained from w by deleting all states of the form (s,t).
2. *For all measurable $\mathfrak{R} \subseteq s_0(S_* \setminus \{s_\bot\})^\omega$ we have*

$$
\mathcal{P}_{\mathcal{M},s_0,\sigma}(\overline{\mathfrak{R}}) \geq \mathtt{val}_{\mathcal{M}}(s_0) \cdot \mathcal{P}_{\mathcal{M}_*,s_0,\sigma}(\mathfrak{R}) \geq \mathcal{P}_{\mathcal{M},s_0,\sigma}(\overline{\mathfrak{R}} \cap [\![\varphi]\!]^{s_0})\,,
$$

   where $\overline{\mathfrak{R}}$ is obtained from $\mathfrak{R}$ by deleting, in all runs, all states of the form (s,t).
3. *We have $\mathtt{val}_{\mathcal{M}}(s_0) \cdot \mathcal{P}_{\mathcal{M}_*,s_0,\sigma}(\varphi) = \mathcal{P}_{\mathcal{M},s_0,\sigma}(\varphi)$. In particular, $\mathtt{val}_{\mathcal{M}_*}(s_0) = 1$, and, for any $\varepsilon \geq 0$, strategy σ is ε-optimal in $\mathcal{M}_*$ if and only if it is $\varepsilon\mathtt{val}_{\mathcal{M}}(s_0)$-optimal in $\mathcal{M}$.*

Lemma 13.3 provides a way of proving the existence of MD strategies that attain, for each state s, a fixed *fraction* (arbitrarily close to 1) of the value of s:

▶ **Theorem 14.** *Let $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P)$ be an MDP, and let φ be an objective that is tail in $\mathcal{M}$. Let $\mathcal{M}_* = (S_*, S_{*\Box}, S_{*\bigcirc}, \longrightarrow_*, P_*)$ be the conditioned version of $\mathcal{M}$ w.r.t. φ. Let $\varepsilon \geq 0$. Any MD strategy σ that is uniformly ε-optimal in $\mathcal{M}_*$ (i.e., $\mathcal{P}_{\mathcal{M}_*,s,\sigma}(\varphi) \geq \mathtt{val}_{\mathcal{M}_*}(s) - \varepsilon$ holds for all $s \in S_*$) is* multiplicatively ε-optimal *in $\mathcal{M}$ (i.e., $\mathcal{P}_{\mathcal{M},s,\sigma}(\varphi) \geq (1-\varepsilon)\mathtt{val}_{\mathcal{M}}(s)$ holds for all $s \in S$).*

Proof. Immediate from Lemma 13.3. ◀

As an application of Theorem 14, we can strengthen the first statement of Theorem 8 towards *multiplicatively* (see Theorem 14) uniform ε-optimal MD strategies for `Transience`.

▶ **Corollary 15.** *In every countable MDP there exist multiplicatively uniform ε-optimal MD strategies for* `Transience`.

Proof. Let $\mathcal{M}$ be a countable MDP, and $\mathcal{M}_*$ its conditioned version w.r.t. `Transience`. Let $\varepsilon > 0$. By Theorem 8, there is a uniform ε-optimal MD strategy σ for `Transience` in $\mathcal{M}_*$. By Theorem 14, strategy σ is multiplicatively uniform ε-optimal in $\mathcal{M}$. ◀

The following lemma, stating that universal transience is closed under "conditioning", is needed for the proof of Lemma 17 below.

▶ **Lemma 16.** *Let $\mathcal{M} = (S, S_\Box, S_\bigcirc, \longrightarrow, P)$ be an MDP, and let φ be an objective that is tail in $\mathcal{M}$. Let $\mathcal{M}_* = (S_*, S_{*\Box}, S_{*\bigcirc}, \longrightarrow_*, P_*)$ be the conditioned version of $\mathcal{M}$ w.r.t. φ, where $s_\perp$ is replaced by an infinite chain $s^1_\perp \longrightarrow s^2_\perp \longrightarrow \cdots$. If $\mathcal{M}$ is universally transient, then so is $\mathcal{M}_*$.*

In [17, Lemma 6] a variant, say $\mathcal{M}_+$, of the conditioned MDP $\mathcal{M}_*$ from Definition 12 was proposed. This variant $\mathcal{M}_+$ differs from $\mathcal{M}_*$ in that $\mathcal{M}_+$ has only those states s from $\mathcal{M}$ that have an optimal strategy, i.e., a strategy σ with $\mathcal{P}_{\mathcal{M},s,\sigma}(\varphi) = \mathtt{val}_{\mathcal{M}}(s)$. Further, for any transition $s \longrightarrow t$ in $\mathcal{M}_+$ where s is a controlled state, we have $\mathtt{val}_{\mathcal{M}}(s) = \mathtt{val}_{\mathcal{M}}(t)$, i.e., $\mathcal{M}_+$ does not have value-decreasing transitions emanating from controlled states. The following lemma was used in the proof of Theorem 10:

▶ **Lemma 17.** *Let $\mathcal{M}$ be an MDP, and let φ be an objective that is tail in $\mathcal{M}$. Let $\mathcal{M}_+$ be the conditioned version w.r.t. φ in the sense of [17, Lemma 6]. If $\mathcal{M}$ is universally transient, then so is $\mathcal{M}_+$.*

7 Conclusion

The `Transience` objective admits ε-optimal (resp. optimal) MD strategies even in *infinitely* branching MDPs. This is unusual, since ε-optimal strategies for most other objectives require infinite memory if the MDP is infinitely branching (in particular all objectives generalizing Safety [17]).

`Transience` encodes a notion of continuous progress, which can be used as a tool to reason about the strategy complexity of other objectives in countable MDPs. E.g., our result on `Transience` is used in [18] as a building block to show upper bounds on the strategy complexity of certain threshold objectives w.r.t. mean payoff, total payoff and point payoff.

References

1 Pieter Abbeel and Andrew 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 Galit Ashkenazi-Golan, János Flesch, Arkadi Predtetchinski, and Eilon Solan. Reachability and safety objectives in Markov decision processes on long but finite horizons. *Journal of Optimization Theory and Applications*, 2020.

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

4 Patrick Billingsley. *Probability and Measure*. Wiley, 1995. Third Edition.

5 Vincent D. Blondel and John N. Tsitsiklis. A survey of computational complexity results in systems and control. *Automatica*, 2000.

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

7 K. Chatterjee and T. Henzinger. A survey of stochastic ω-regular games. *Journal of Computer and System Sciences*, 2012.

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

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

10 William Feller. *An Introduction to Probability Theory and Its Applications*. Wiley & Sons, second edition, 1966.

11 János Flesch, Arkadi Predtetchinski, and William Sudderth. Simplifying optimal strategies in limsup and liminf stochastic games. *Discrete Applied Mathematics*, 2018.

12 T.P. Hill and V.C. Pestien. The existence of good Markov strategies for decision processes with general payoffs. *Stoch. Processes and Appl.*, 1987.

13 S. Kiefer, R. Mayr, M. Shirmohammadi, and P. Totzke. Transience in countable MDPs. In *International Conference on Concurrency Theory*, LIPIcs, 2021. Full version at arXiv:2012.13739.

14 Stefan Kiefer, Richard Mayr, Mahsa Shirmohammadi, and Patrick Totzke. Büchi objectives in countable MDPs. In *International Colloquium on Automata, Languages and Programming*, LIPIcs, 2019. Full version at arXiv:1904.11573. doi:10.4230/LIPIcs.ICALP.2019.119.

15 Stefan Kiefer, Richard Mayr, Mahsa Shirmohammadi, and Patrick Totzke. Strategy Complexity of Parity Objectives in Countable MDPs. In *International Conference on Concurrency Theory*, 2020. doi:10.4230/LIPIcs.CONCUR.2020.7.

16 Stefan Kiefer, Richard Mayr, Mahsa Shirmohammadi, Patrick Totzke, and Dominik Wojtczak. How to play in infinite MDPs (invited talk). In *International Colloquium on Automata, Languages and Programming*, 2020. doi:10.4230/LIPIcs.ICALP.2020.3.

17 Stefan Kiefer, Richard Mayr, Mahsa Shirmohammadi, and Dominik Wojtczak. Parity Objectives in Countable MDPs. In *Annual IEEE Symposium on Logic in Computer Science*, 2017. doi:10.1109/LICS.2017.8005100.

18 Richard Mayr and Eric Munday. Strategy Complexity of Mean Payoff, Total Payoff and Point Payoff Objectives in Countable MDPs. In *International Conference on Concurrency Theory*, LIPIcs, 2021. The full version is available at arXiv:2107.03287.

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

20 Donald Ornstein. On the existence of stationary optimal strategies. *Proceedings of the American Mathematical Society*, 1969. doi:10.2307/2035700.

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

22 George Santayana. Reason in common sense, 1905. In *Volume 1 of The Life of Reason*. URL: https://en.wikipedia.org/wiki/George_Santayana.

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

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

25 William D. Sudderth. Optimal Markov strategies. *Decisions in Economics and Finance*, 2020.

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

27 Moshe Y. Vardi. Automatic verification of probabilistic concurrent finite-state programs. In *Annual Symposium on Foundations of Computer Science*. IEEE Computer Society, 1985. doi:10.1109/SFCS.1985.12.