Reachability and Safety Objectives in Markov Decision Processes on Long but Finite Horizons

Galit Ashkenazi-Golan, János Flesch, Arkadi Predtetchinski & Eilon Solan

Journal of Optimization Theory and Applications

ISSN 0022-3239

J Optim Theory Appl
DOI 10.1007/s10957-020-01681-2

Springer

Journal of Optimization Theory and Applications
https://doi.org/10.1007/s10957-020-01681-2

Reachability and Safety Objectives in Markov Decision Processes on Long but Finite Horizons

Galit Ashkenazi-Golan[1] · János Flesch[2] · Arkadi Predtetchinski[3] · Eilon Solan[1]

Received: 11 November 2019 / Accepted: 28 April 2020

Abstract
We consider discrete-time Markov decision processes in which the decision maker is interested in long but finite horizons. First we consider reachability objective: the decision maker's goal is to reach a specific target state with the highest possible probability. A strategy is said to overtake another strategy, if it gives a strictly higher probability of reaching the target state on all sufficiently large but finite horizons. We prove that there exists a pure stationary strategy that is not overtaken by any pure strategy nor by any stationary strategy, under some condition on the transition structure and respectively under genericity. A strategy that is not overtaken by any other strategy, called an overtaking optimal strategy, does not always exist. We provide sufficient conditions for its existence. Next we consider safety objective: the decision maker's goal is to avoid a specific state with the highest possible probability. We argue that the results proven for reachability objective extend to this model.

Communicated by Jörg Rambau.

✉ János Flesch
j.flesch@maastrichtuniversity.nl

Galit Ashkenazi-Golan
galit.ashkenazi@gmail.com

Arkadi Predtetchinski
a.predtetchinski@maastrichtuniversity.nl

Eilon Solan
eilons@post.tau.ac.il

[1] School of Mathematical Sciences, Tel-Aviv University, Tel Aviv, Israel

[2] Department of Quantitative Economics, Maastricht University, Maastricht, The Netherlands

[3] Department of Economics, Maastricht University, Maastricht, The Netherlands

Published online: 18 May 2020

Keywords Markov decision process · Reachability objective · Safety objective · Overtaking optimality · Perron–Frobenius eigenvalue

Mathematics Subject Classification 90C40

1 Introduction

We consider discrete-time Markov decision processes (MDP) with finite state and action spaces. We consider two different types of objectives for the decision maker: reachability objectives and safety objectives. The decision maker is said to have reachability objective, if his goal is to reach a specific state of the MDP with the highest possible probability, and the decision maker is said to have a safety objective, if his goal is the opposite: to avoid a specific state of the MDP with the highest possible probability. Both objectives are standard and have been analyzed extensively in the literature, but they are quite different in nature (see, e.g., [1–3]).

An important question is on which time horizon the decision maker evaluates his strategies. On any given finite horizon, backward induction guarantees that the decision maker has a pure optimal strategy. This optimal strategy can depend heavily on the horizon, and generally there is no strategy that is optimal on all finite horizons. On the infinite horizon, the decision maker has a pure stationary optimal strategy (cf. [4,5]).

In this paper, instead of considering a fixed horizon, we propose to evaluate strategies by how they perform on all long but finite horizons. In particular, such an evaluation can be meaningful, if the decision maker knows that the decision process will last for many periods, but he has no information on its exact length. In the case of reachability objectives, such an evaluation may also reflect the attitude of a decision maker who is patient and can wait for many periods to reach the target state.

More precisely, when the decision maker has reachability objective with target state s^*, we say that a strategy σ overtakes another strategy σ', if there exists $T \in \mathbb{N}$ such that, on all finite horizons $t \geq T$, the probability of having visited the state s^* within horizon t is strictly larger under σ than under σ'. Thus, conditionally on the MDP lasting at least T periods, σ performs better than σ' regardless of the horizon, and consequently the decision maker should prefer σ to σ'. When the decision maker has a safety objective and wants to avoid a state s^*, we say that a strategy σ overtakes another strategy σ', if there exists $T \in \mathbb{N}$ such that, on all finite horizons $t \geq T$, the probability of having visited the state s^* within horizon t is strictly smaller under σ than under σ'.

We also define a more permissive version of the aforementioned relations between strategies. For reachability objectives, we say that a strategy σ weakly overtakes another strategy σ', if there exists $T \in \mathbb{N}$ such that, on all finite horizons $t \geq T$, the probability of having visited the target state s^* within horizon t under σ is at least as much as that under σ', but strictly more for infinitely many horizons t. The definition is analogous for safety objectives.

Under these comparisons of strategies, we call a strategy overtaking optimal, if it is not overtaken by any other strategy and call it strongly overtaking optimal, if it is not weakly overtaken by any other strategy. Strong overtaking optimality is a strict

refinement of overtaking optimality, and as an appealing property, they are both strict refinements of optimality on the infinite horizon.

Our contribution. For reachability objectives, we obtain the following results, sorted by the attributes of the MDP. (I.1) We prove that if the MDP is such that each action can lead to at most one non-target state with a positive probability, then there exists a pure stationary strategy that is not weakly overtaken by any pure strategy. This is Theorem 4.1. We show with Example 4.2 that such a statement does not hold for all MDPs. (I.2) We prove by means of Example 4.1 that an overtaking optimal strategy does not always exist. This MDP is however constructed in a very specific way and is non-generic. (I.3) We consider MDPs that are generic, in the sense that the transition probabilities are randomized using any non-trivial joint density function. We show for these MDPs that there exists a pure stationary strategy that overtakes each other stationary strategy. This is Theorem 5.1. (I.4) We present sufficient conditions in Theorem 6.1 for the existence of a stationary strategy that is strongly overtaking optimal. For safety objectives, we argue that the same results hold.

Proof techniques. We use quite different proof techniques to obtain our results. For proving result (I.1), we transform the MDP with the reachability objective into a regular MDP, by assigning payoffs to actions based on the immediate transition probabilities to the target state. In this new MDP, we invoke some results in [6] to derive a specific pure stationary strategy. We show that this strategy is exactly the desired strategy in the original MDP with the reachability objective. This proof technique is suitable for pure strategies, but probably also limited to them, as the relation between the two MDPs is much weaker for non-pure strategies. When considering generic MDPs in result (I.3), we rely on techniques from linear algebra. The overtaking comparison between two stationary strategies can be reduced to the comparison of the spectral gaps of the transition matrices that these strategies induce. The spectral gap of a transition matrix refers to the difference between the largest eigenvalue, which is equal to 1, and the modulus of the second eigenvalue, which can be a complex number. To obtain result (I.3), we need to compare the spectral gaps of transition matrices induced by stationary and pure stationary strategies. Result (I.4) is proven in a constructive way. The mixed actions of the desired stationary strategy can be derived from the conditions of Theorem 6.1. The results for safety objectives are proven similarly.

Related literature. Reachability and safety problems were studied both in the MDP framework and in the context of two-player zero-sum games, for an overview we refer to [1] and respectively to [2] and [3]. An important distinction is made between the qualitative and the quantitative approaches. The qualitative approach is interested in the probability with which the decision maker succeeds to meet his objective. For the quantitative approach, however, it also matters how quickly the target state is reached in the case of reachability objective, or how long the bad state has been avoided in the case of a safety objective. Our overtaking approach could thus be classified as a quantitative approach on the infinite horizon. For other quantitative approaches, we refer to [7,8] and the references therein, and to [9].

In the literature, various definitions of overtaking optimality have been proposed. They all serve as a refinement of optimality on the infinite horizon, based on the performance of strategies on the finite horizons. For an overview, we refer to [10–16].

A well-established definition of overtaking optimality is given in Section 5.4.2 in [11] for MDPs in which the decision maker receives a payoff at each period, depending on the state and the chosen action. According to this definition, a strategy σ^* is overtaking optimal if, for each strategy σ and for each error-term $\delta > 0$ the following holds: for all large horizons N, the expected sum of the payoffs during the first N periods under σ^* is at least as much as that under σ minus δ. In our framework, there are no immediate payoffs, hence this definition does not apply. However, we show in Example 3.1 that, if we take the natural assignment of payoff 0 to each non-target state and payoff 1 to the target state, then our definition of overtaking optimality can lead to different strategies than Puterman's definition, as well as variants of Puterman's definition defined therein.

Our definition of overtaking optimality is a relatively direct translation of the definitions of sporadic overtaking optimality in [6,10] and repeated optimality in [16], into the context of MDPs with reachability and safety objectives.

One important feature of our definition is that it does not require the strategy to outperform all other strategies on long but finite horizons. It only requires that the strategy is not outperformed by any other strategy. Our definition is therefore weaker than overtaking optimality and uniform overtaking optimality as in [10], and weaker than strong overtaking optimality as in [17] or [18]. See also [16], who delineates "not-outperformed" definitions from "outperform-all" definitions of optimality.

Organization of the paper. Section 2 details the model. Then, we start by analyzing reachability objectives. Section 3 provides an example, which highlights different aspects of the concept of overtaking optimality by comparing it with other optimality notions. Sections 4 and 5 present the results for piecewise deterministic MDPs and generic MDPs, respectively. Section 6 provides sufficient conditions that ensure that a stationary strategy is strongly overtaking optimal. In Sect. 7, we turn to safety objectives. Section 8 concludes.

2 The Model

2.1 MDPs with Reachability Objective

The model. An *MDP* is given by [1] a nonempty, finite set S of states, [2] for each state $s \in S$, a nonempty, finite set $A(s)$ of actions and [3] for each state $s \in S$ and action $a \in A(s)$, a probability distribution $p(s,a) = (p(z \mid s,a))_{z \in S}$ on the set S of states. The MDP is played at periods in $\mathbb{N} = \{1, 2, \ldots\}$ as follows: The initial state s_1 is given. At each period t, the decision maker chooses an action $a_t \in A(s_t)$, which leads to a state $s_{t+1} \in S$ drawn according to the distribution $p(s_t, a_t)$. An *MDP with reachability objective* is an MDP together with a specific state $s^* \in S$, called the *target state*, which is not the initial state s_1.

Histories. Let H_∞ be the set of all *infinite histories*, i.e., the set of sequences $(s_1, a_1, s_2, a_2, \ldots)$ such that $s_i \in S$, $a_i \in A(s_i)$, and $p(s_{i+1} \mid s_i, a_i) > 0$ for each $i \in \mathbb{N}$. A *history at period t* is a prefix $(s_1, a_1, \ldots, s_{t-1}, a_{t-1}, s_t)$ of an infinite history.

Denote by H_t the set of all histories at period t, by $H = \cup_{t\in\mathbb{N}} H_t$ the set of all histories, and by $s(h)$ the final state of each history $h \in H$.

Strategies. A *mixed action* in a state $s \in S$ is a probability distribution on $A(s)$. The set of mixed actions in state s is denoted by $\Delta(A(s))$. A *strategy* σ is a map that, to each history $h \in H$, assigns a mixed action $\sigma(h) \in \Delta(A(s(h)))$. The interpretation is that, if history h arises, σ chooses an action according to the probabilities given by the mixed action $\sigma(h)$. A strategy σ is called *pure*, if $\sigma(h)$ places probability 1 on one action, for each history h. A strategy σ is called *stationary*, if the recommendation of the action only depends on the current state, i.e., $\sigma(h) = \sigma(h')$ whenever $s(h) = s(h')$. Note that a pure stationary strategy can be seen as an element of $\times_{s\in S} A(s)$. Every initial state s and strategy σ induce a probability measure $\mathbb{P}_{s\sigma}$ on H_∞, where H_∞ is endowed with the sigma–algebra generated by the cylinder sets. We denote the corresponding expectation operator by $\mathbb{E}_{s\sigma}$.

Value and optimality. Let t^* denote the first period when state s^* is reached; if s^* is not reached then $t^* = \infty$. The *value* at the initial state s is the maximal probability that state s^* can be reached: $v(s) = \sup_\sigma \mathbb{P}_{s\sigma}(t^* < \infty)$. A strategy σ is called *optimal* at the initial state s if $\mathbb{P}_{s\sigma}(t^* < \infty) = v(s)$. It is known that the decision maker always has a pure stationary strategy that is optimal at all initial states (cf. [4,5]).

Overtaking optimality. We say that a strategy σ *overtakes* a strategy σ' at the initial state s if there is $T \in \mathbb{N}$ such that for all periods $t \geq T$ we have $\mathbb{P}_{s\sigma}(t^* \leq t) > \mathbb{P}_{s\sigma'}(t^* \leq t)$. This means that, for all periods $t \geq T$, the probability under σ to reach s^* within the first t periods is strictly larger than that under σ'. If the decision maker is sufficiently patient with regard to his goal to reach the target state, then he strictly prefers σ to σ'.

Note that two strategies can be incomparable in the sense that neither of them overtakes the other one. Consider the following example. The state space is $\{x, s^*\}$. In state x, the decision maker has three actions: $a_0, a_{1/2}, a_{7/8}$. For $z \in \{0, 1/2, 7/8\}$, under action a_z, the play moves to state s^* with probability z and remains in state x with probability $1 - z$. Now suppose that σ recommends to always play action $a_{1/2}$ and σ' recommends to play the sequence of actions $a_0, a_{7/8}, a_0, a_0, a_{7/8}, a_0, \ldots$ as long as the play is in state x. Then, at periods $t = 3k + 1$, where $k \in \mathbb{N}$, we have $\mathbb{P}_{s\sigma}(t^* \leq t) = \mathbb{P}_{s\sigma'}(t^* \leq t) = (7/8)^k$. At periods $t = 3k + 2$, we have $\mathbb{P}_{s\sigma}(t^* \leq t) > \mathbb{P}_{s\sigma'}(t^* \leq t)$. At periods $t = 3k$, we have $\mathbb{P}_{s\sigma}(t^* \leq t) < \mathbb{P}_{s\sigma'}(t^* \leq t)$. So, σ and σ' are incomparable.

A strategy σ is called *overtaking optimal* at the initial state s, if there is no strategy that overtakes σ at that initial state. That is, σ is maximal with respect to the relation of "overtakes" between strategies.

Note that any optimal strategy overtakes any strategy that is not optimal. Indeed, if strategy σ is optimal at the initial state s but strategy σ' is not, then $\mathbb{P}_{s\sigma}(t^* < \infty) = v(s) > \mathbb{P}_{s\sigma'}(t^* < \infty)$, and hence $\mathbb{P}_{s\sigma}(t^* \leq t) > \mathbb{P}_{s\sigma'}(t^* \leq t)$ for all sufficiently large t. Consequently, an overtaking optimal strategy at the initial state s is also optimal at that initial state. As Example 3.1 will show, the converse is not true: there exist optimal strategies at an initial state that are not overtaking optimal at that initial state. Thus, overtaking optimality is a strict refinement of optimality.

Strong overtaking optimality. We say that a strategy σ *weakly overtakes* a strategy σ' at the initial state s if there is $T \in \mathbb{N}$ such that for all periods $t \geq T$ we have $\mathbb{P}_{s\sigma}(t^* \leq t) \geq \mathbb{P}_{s\sigma'}(t^* \leq t)$ with strict inequality for infinitely many t. Note that if σ overtakes σ' at the initial state s then σ also weakly overtakes σ' at that initial state.

A strategy σ is called *strongly overtaking optimal* at the initial state s, if no strategy weakly overtakes σ at that initial state. A strongly overtaking optimal strategy at an initial state is also overtaking optimal at that initial state.

2.2 Discounted and Average Payoff MDPs

We will also consider MDPs with the discounted payoff or with the average payoff, but only as auxiliary models. A *discounted MDP* is an MDP together with a discount factor $\beta \in]0, 1[$ and a payoff function, namely a function $(s, a) \mapsto u(s, a) \in \mathbb{R}$ that maps a payoff to each state $s \in S$ and action $a \in A(s)$. For initial state $s \in S$, the β-discounted value is defined as

$$v_\beta(s) = \sup_\sigma \mathbb{E}_{s\sigma}\left[(1-\beta)\cdot\sum_{t=1}^{\infty}\beta^{t-1}\cdot u(s_t, a_t)\right]. \tag{1}$$

A strategy σ is called β-discounted optimal at the initial state s, if the supremum in (1) is attained at σ. A strategy σ^* is called *Blackwell optimal*, if there is $B \in]0, 1[$ such that σ^* is β-discounted optimal at all initial states for all discount factors $\beta \in [B, 1[$.

By the results of [19] and [5], it is known that for each discount factor $\beta \in]0, 1[$, the decision maker has a pure stationary strategy that is β-discounted optimal at all initial states, and that he has a Blackwell optimal strategy too.

An *average payoff MDP* is similar to a discounted MDP, except that the decision maker's goal is to maximize the expectation of the average payoff $\liminf_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} u(s_t, a_t)$. The average value and average optimality are defined analogously to the corresponding definitions for reachability objective. By [4,5], it is known that the decision maker has a pure stationary strategy that is average optimal at all initial states. Moreover, each Blackwell optimal strategy is average optimal at all initial states.

3 Reachability Objectives: An Illustrative Example

In this section, we discuss a specific MDP with reachability objective, which demonstrates four properties of overtaking optimality. First, there are optimal strategies that are not overtaking optimal. That is, overtaking optimality is a strict refinement of optimality. Second, finding overtaking optimal strategies cannot be done by simply solving a related discounted MDP. Third, the strategy that minimizes the expected time of reaching the target state s^* can be different from the overtaking optimal strategies, even when the latter is unique. Fourth, for a related MDP, overtaking optimality in not the same as the concept of overtaking optimality as defined in [11], or the related

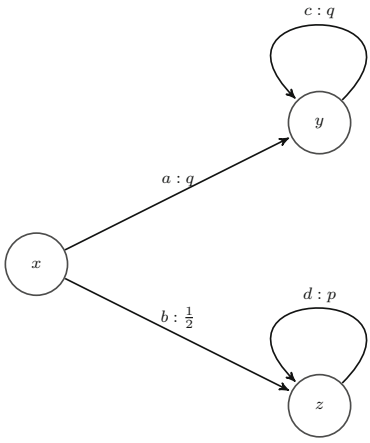

Fig. 1 The MDP in Example 3.1

concepts of cumulative overtaking optimality and average overtaking optimality that were defined therein.

Example 3.1 Consider an MDP that has state space $S = \{x, y, z, s^*\}$ such that:

- State x is the initial state. In this state, the decision maker has two actions: a and b. Action a leads to state s^* with probability q and to state y with probability $1 - q$. Action b leads to state s^* with probability $\frac{1}{2}$ and to state z with probability $\frac{1}{2}$.
- In state y, there is only one action, denoted by c, which leads to state s^* with probability q and to state y with probability $1 - q$.
- In state z, there is only one action, denoted by d, which leads to state s^* with probability p and to state z with probability $1 - p$.
- State s^* is absorbing.

The probabilities p and q are such that $0 < p < q < \frac{2p}{2p+1}$. For example, $p = 0.1$ and $q = 0.11$. Note that $p < \frac{2p}{2p+1}$ implies $p < 1/2$, and therefore $p < q < \frac{2p}{2p+1}$ implies $q < 1/2$ as well.

The MDP is depicted in Fig. 1. In this figure, the state s^* is omitted for simplicity, states x, y, and z are denoted by circles, and actions are denoted by arcs together with the name of the action and the corresponding probability of moving to state s^*. For example, the arrow from state x to state y denoted by $a : q$ indicates that this action is called a and that, when played at state x, it leads to state s^* with probability q and to state y with probability $1 - q$.

Suppose that the initial state is x. Since in states y and z there is a single action, a strategy is characterized by the action it selects in state x. Choosing the action a at state x leads to the following sequence of probabilities of moving to state s^*: $(q, q, q, \ldots)$. Choosing the action b at state x leads to the following sequence of probabilities of moving to state s^*: $(1/2, p, p, \ldots)$. The state s^* is eventually reached with probability 1 under both actions a and b. Below we argue that while b (and not a) is optimal according to many optimality concepts, a (and not b) is overtaking optimal under reachability objective.

Claim 1 In Example 3.1, action a is overtaking optimal under reachability objective, and action b is not.

Proof Take a period $t \geq 2$. Under a the probability of reaching the target state s^* within the first t periods is $\mathbb{P}_{xa}(t^* \leq t) = 1 - (1-q)^{t-1}$, whereas under b this probability is $\mathbb{P}_{xb}(t^* \leq t) = \frac{1}{2} + \frac{1}{2} \cdot \left(1 - (1-p)^{t-2}\right)$. Thus,

$$\begin{aligned}\mathbb{P}_{xa}(t^* \leq t) - \mathbb{P}_{xb}(t^* \leq t) &= -(1-q)^{t-1} + \frac{1}{2} \cdot (1-p)^{t-2} \\ &= (1-p)^{t-2} \cdot \left[-\left(\frac{1-q}{1-p}\right)^{t-2} \cdot (1-q) + \frac{1}{2} \right].\end{aligned}$$

Since $p < q$ by assumption, $(1-q)/(1-p) < 1$. So, $\mathbb{P}_{xa}(t^* \leq t) - \mathbb{P}_{xb}(t^* \leq t)$ is positive for large $t \in \mathbb{N}$. This completes the proof. □

Claim 2 Consider the discounted MDP that has payoff equal to 1 in state s^* and payoff 0 in states x, y and z. Strategy b is Blackwell optimal, while strategy a is not Blackwell optimal.

Proof In the discounted MDP, if state s^* is reached at some period t, then the β-discounted payoff is equal to $(1-\beta) \cdot (\beta^{t-1} + \beta^t + \cdots) = \beta^{t-1}$. Thus, action a leads to the expected discounted payoff $D_a(\beta) = q\beta \frac{1}{1-(1-q)\beta}$, whereas action b leads to the expected discounted payoff: $D_b(\beta) = \frac{1}{2}\beta + \frac{1}{2} p\beta^2 \frac{1}{1-(1-p)\beta}$. As one can verify, we have

$$D_b(\beta) - D_a(\beta) = \frac{\frac{1}{2}\beta(1-\beta) \cdot \left(1 - (1-2p)(1-q)\beta - 2q\right)}{(1-(1-p)\beta) \cdot (1-(1-q)\beta)}.$$

The denominator of the fraction above is positive for all $\beta \in]0, 1[$. As $q < \frac{2p}{2p+1}$ by assumption, the expression $1-(1-2p)(1-q)-2q$ is positive. Hence, the numerator of the fraction above is positive for large $\beta \in]0, 1[$. Thus, we find $D_b(\beta) > D_a(\beta)$ for large $\beta \in]0, 1[$, and the claim follows. □

Claim 3 In Example 3.1, the expectation of the period t^* when reaching the state s^* is smaller under b than under a: $\mathbb{E}_{xb}(t^*) < \mathbb{E}_{xa}(t^*)$.

Proof We have $\mathbb{E}_{xa}(t^*) = 2q + 3(1-q)q + 4(1-q)^2 q + \cdots = \frac{1}{q} + 1$ and $\mathbb{E}_{xb}(t^*) = 2 \cdot \frac{1}{2} + \frac{1}{2} \cdot [3p + 4(1-p)p + 5(1-p)^2 p + \cdots] = \frac{1}{2p} + 2$. Since $q < \frac{2p}{2p+1}$ by assumption, the claim follows. □

Claim 4 Consider the MDP with payoff 1 in state s^* and payoff 0 in states x, y, and z. The strategy b is both overtaking optimal and cumulative overtaking optimal according to the definitions in Section 5.4.2 in [11], and strategy a is neither of them.[1]

[1] Both a and b are average overtaking optimal according to Puterman's definition.

Proof For each strategy σ and period N, define $R(\sigma, N)$ to be the expected sum of the payoffs (which is also the sum of the expected payoffs) up to period N. Since the initial state x is not the target state, $R(a, 1) = R(b, 1) = 0$ holds trivially. We claim that $R(a, N) < R(b, N)$ for every $N \geq 2$, and, moreover, $\lim_{N\to\infty} R(b, N) - R(a, N) = -\frac{1}{2p} + \frac{1-q}{q} > 0$.

Since $\mathbb{P}_{xa}(t^* \leq t) = 1 - (1-q)^{t-1}$, for every $N \geq 2$ we have

$$R(a, N) = \sum_{t=2}^{N} \left(1 - (1-q)^{t-1}\right) = (N-1) - (1-q) \cdot \frac{1 - (1-q)^{N-1}}{1 - (1-q)},$$

and since $\mathbb{P}_{xb}(t^* \leq t) = \frac{1}{2} + \frac{1}{2} \cdot \left(1 - (1-p)^{t-2}\right)$, we have

$$R(b, N) = \sum_{t=2}^{N} \left(\frac{1}{2} + \frac{1}{2} \cdot (1 - (1-p)^{t-2})\right) = (N-1) - \frac{1}{2} \cdot \frac{1 - (1-p)^{N-1}}{1 - (1-p)}.$$

Since $p < q$ and $q < \frac{2p}{2p+1}$ by assumption, $1 - (1-q)^{N-1} > 1 - (1-p)^{N-1}$ and $\frac{1-q}{q} > \frac{1}{2p}$, and therefore $(1-q) \cdot \frac{1-(1-q)^{N-1}}{1-(1-q)} > \frac{1}{2} \cdot \frac{1-(1-p)^{N-1}}{1-(1-p)}$. Thus, $R(a, N) < R(b, N)$, and therefore b (and not a) is cumulative overtaking optimal according to Puterman's definition. Moreover,

$$\lim_{N\to\infty} R(b, N) - R(a, N) = -\frac{1}{2p} + \frac{1-q}{q} > 0,$$

so b (and not a) is overtaking optimal according to Puterman's definition. □

4 Reachability Objectives: Piecewise Deterministic MDPs

A *piecewise deterministic Markov process* [20] is a process whose behavior is governed by random jumps at points in time, but whose evolution is deterministically governed by an ordinary differential equation between those times. These processes have been shown to be useful in a wide range of applications, including queueing theory, ruin problems, biochemistry and geology. In this section, we study an analogous concept when the state space is finite and when the jumps lead to the target state.

We call an MDP with reachability objective *piecewise deterministic* if for each state $s \neq s^*$ and each action $a \in A(s)$ there is a state $\omega(s, a) \in S$ such that $p(\{\omega(s, a), s^*\} \mid s, a) = 1$. That is, for any state and action, the play moves to the target state s^* or to a specific state that is not s^*.

A special case of piecewise deterministic MDPs is that of deterministic MDPs (see Section 3.3 in [11]), that is, when there is no randomness in the transitions: for every state $s \in S$ and action $a \in A(s)$ there is a unique state $w(s, a) \in S$ such that $p(w(s, a) \mid s, a) = 1$.

The following theorem states that, in piecewise deterministic MDPs with reachability objective, there always exists a pure stationary strategy that is at least as good as

any pure strategy in the overtaking sense. The main idea of the proof is to transform the MDP with reachability objective into an average payoff MDP. The payoffs that we assign to actions are related to the probabilities that these actions lead to the target state.

The condition that the MDP is piecewise deterministic plays an important role. Indeed, Example 4.2 will show that the result is not true in general for MDPs that are not piecewise deterministic.

Theorem 4.1 *In every piecewise deterministic* MDP *with reachability objective, there exists a pure stationary strategy that is not weakly overtaken by any other pure strategy.*

To prove Theorem 4.1, we need the following result. The proof is provided in the Appendix.

Theorem 4.2 *Consider a deterministic discounted* MDP *and let σ be a Blackwell optimal strategy. There exists no pure strategy σ' and no initial state $s \in S$ with the following properties:*

- *Property I: there is $M \in \mathbb{N}$ such that for all periods $t \geq M$ we have $u_t(s,\sigma) \leq u_t(s,\sigma')$, where $u_t(s,\sigma)$ and $u_t(s,\sigma')$ are the expected average payoffs up to period t under σ and respectively under σ' at initial state s,*
- *Property II: $u_t(s,\sigma) < u_t(s,\sigma')$ holds for infinitely many t.*

Proof of Theorem 4.1 Consider a piecewise deterministic MDP $\mathcal{M}$ with reachability objective. We may assume[2] that there is no state $s \neq s^*$ and action $a \in A(s)$ such that $p(s^* \mid s,a) = 1$. As the MDP $\mathcal{M}$ is piecewise deterministic, this implies that $p(\omega(s,a) \mid s,a) > 0$, for every $s \in S$ and $a \in A(s)$.

We define an auxiliary average payoff deterministic MDP $\mathcal{M}'$ as follows: (i) The state space is $S' = S - \{s^*\}$. (ii) For each state $s \in S'$, the action space is the same as in the MDP $\mathcal{M}$ with reachability objective: $A'(s) = A(s)$. (iii) For each state $s \in S'$ and action $a \in A(s)$, the transition and the payoff in $\mathcal{M}'$ are defined as follows: $p'(w(s,a) \mid s,a) = 1$ and $u'(s,a) = -\log(p(w(s,a) \mid s,a))$.

Intuitively, the MDP $\mathcal{M}'$ represents what happens if in the MDP $\mathcal{M}$ the decision maker is unlucky at all periods and the process never reaches state s^*. Let σ be a pure Blackwell optimal stationary strategy in $\mathcal{M}'$. We show that σ is not weakly overtaken by any pure strategy, thereby proving the theorem.

Consider any other pure strategy ρ in $\mathcal{M}'$. Since the MDP $\mathcal{M}'$ is deterministic, each of the pure strategies σ and ρ induces a specific infinite history in $\mathcal{M}'$ with probability 1. Let $(s_t,a_t)_{t\in\mathbb{N}}$ denote the infinite history induced by σ and $(z_t,b_t)_{t\in\mathbb{N}}$ denote the infinite history induced by ρ.

In the original MDP $\mathcal{M}$, the probability under σ that state s^* is not reached within the first t periods is $1 - \mathbb{P}_{s_1,\sigma}(t^* \leq t)$. This probability is related to the payoffs in the average payoff MDP $\mathcal{M}'$. Indeed, for each period $t \geq 2$ we have

$$\log\left(1 - \mathbb{P}_{s_1,\sigma}(t^* \leq t)\right) = \log\left(\prod_{k=1}^{t-1} p(s_{k+1} \mid s_k,a_k)\right) = \sum_{k=1}^{t-1} \log\left(p(s_{k+1} \mid s_k,a_k)\right)$$

[2] Indeed, in such a state s it is optimal to choose such an action a, as it leads in one step to state s^*. Hence, such a state can be deleted from the MDP, and each transition to state s can be rewritten as a transition directly to s^*.

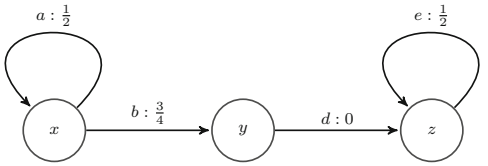

Fig. 2 The MDP in Example 4.1

$$= -\sum_{k=1}^{t-1} u'(s_k, a_k) = -(t-1) \cdot u_{t-1}(s_1, \sigma).$$

Similarly, $\log\left(1 - \mathbb{P}_{s_1,\rho}(t^* \leq t)\right) = -\sum_{k=1}^{t-1} u'(z_k, b_k) = -(t-1) \cdot u_{t-1}(s_1, \rho)$.

By the choice of σ in $\mathcal{M}'$ (cf. Theorem 4.2), one of the following holds: (i) There is $M \in \mathbb{N}$ such that for all periods $t \geq M$ we have $u_t(s_1, \sigma) = u_t(s_1, \rho)$. (ii) There is a strictly increasing sequence $(t_k)_{k\in\mathbb{N}}$ of periods such that for each $k \in \mathbb{N}$ we have $u_{t_k}(s_1, \sigma) > u_{t_k}(s_1, \rho)$.

If (i) holds, then we have $\mathbb{P}_{s_1,\sigma}(t^* \leq t) = \mathbb{P}_{s_1,\rho}(t^* \leq t)$ for all $t \geq M + 1$, and hence ρ does not weakly overtake σ in the original MDP $\mathcal{M}$. If (ii) holds, then $\mathbb{P}_{s_1,\sigma}(t^* \leq t_k) > \mathbb{P}_{s_1,\rho}(t^* \leq t_k)$ for each $k \in \mathbb{N}$, and hence ρ does not weakly overtake σ in the original MDP $\mathcal{M}$ in this case either. □

The following example demonstrates that, even if the MDP with reachability objective is piecewise deterministic, an overtaking optimal strategy may fail to exist. In the example, each pure strategy is equally good in the overtaking sense, but each pure strategy is overtaken by any strategy that uses randomization at every period.

Example 4.1 Consider the MDP with reachability objective given in Fig. 2, with a notation similar to that of Example 3.1. The initial state is state x.

Since in states y and z there is a single action, a pure strategy is characterized by the period in which the action b is first played in state x. When playing action b and subsequently action d, the total probability during these two periods of reaching the target state is $\frac{3}{4}$. Playing action a twice (or action e twice) leads to the same total probability, as $\frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{4}$. It follows that $\mathbb{P}_{x\sigma}(t^* \leq t) = 1 - \frac{1}{2^{t-1}}$, for all pure strategies σ, except of the strategy $\sigma' = a^{t-2}b$ that plays the action a in the first $t-2$ periods and the action b in period $t-1$, for which $\mathbb{P}_{x\sigma'}(t^* \leq t) > 1 - \frac{1}{2^{t-1}}$. This implies that strategies that use randomization at state x at every period do better in the overtaking sense than all pure strategies. Indeed, given any $t \geq 2$, when calculating the probability of reaching the target state within the first t periods, there is a positive probability that action b is played exactly at period $t-1$ (and thus we reach the target state exactly at period t), while not having to include consequences of playing action d yet.

Claim 1 Consider Example 4.1. For each pure strategy σ, it holds for sufficiently large periods t that the probability of reaching the target state within the first t periods is $\mathbb{P}_{x\sigma}(t^* \leq t) = 1 - \frac{1}{2^{t-1}}$. In particular, no pure strategy is overtaken by another pure strategy.

Proof For the pure strategy a^∞ that plays a at all periods, we have for all periods t that $\mathbb{P}_{xa^\infty}(t^* \leq t) = 1 - \frac{1}{2^{t-1}}$. For any other pure strategy $a^{n-1}b$ that plays a at the first $n-1$ periods and plays b at period n, we have for all periods $t \geq n+1$ that $\mathbb{P}_{x,a^{n-1}b}(t^* \leq t) = 1 - \frac{1}{2^{t-1}}$. □

Claim 2 Consider Example 4.1. Take two strategies σ and σ'. Consider a period $t \geq 2$. Then, $\mathbb{P}_{x\sigma}(t^* \leq t) > \mathbb{P}_{x\sigma'}(t^* \leq t)$ holds if and only if the probability under σ of being in state x and playing action b at period $t-1$ is strictly larger than that under σ', i.e., $\mathbb{P}_{x\sigma}(a_{t-1} = b) > \mathbb{P}_{x\sigma'}(a_{t-1} = b)$. Consequently, if the condition $\mathbb{P}_{x\sigma}(a_{t-1} = b) > \mathbb{P}_{x\sigma'}(a_{t-1} = b)$ holds for all sufficiently large periods t, then σ overtakes σ'.

Proof Suppose that when playing two strategies σ and σ', it holds that for some period $t \geq 2$ we have $\mathbb{P}_{x\sigma}(a_{t-1} = b) > \mathbb{P}_{x\sigma'}(a_{t-1} = b)$.

On the finite horizon up to period t, the set of pure strategies is the finite set $W^t = \{a^t,\ b,\ ab,\ a^2b, \ldots,\ a^{t-1}b\}$. Under $a^{t-2}b$, the probability of reaching the target state within the first t periods is $\mathbb{P}_{x,a^{t-2}b}(t^* \leq t) = 1 - \frac{1}{2^{t-2}} \cdot \frac{1}{4} = 1 - \frac{1}{2^t}$, while under each other pure strategy $\tau \neq a^{t-2}b$, this is $\mathbb{P}_{x\tau}(t^* \leq t) = 1 - \frac{1}{2^{t-1}}$.

The strategy σ induces in a natural way a probability distribution on the finite set W^t of pure strategies. Indeed, denote by $((x,a)^{k-1}, x)$ the history at period k in which the play remained through action a in state x until period k, and by $\sigma(h)(a)$ the probability to select action a after history h. Then, we have: (i) $\mathbb{P}_{x\sigma}(b) = \sigma(x)(b)$, since (x) is the history at period 1. (ii) $\mathbb{P}_{x\sigma}(a^t) = \sigma(x)(a) \cdot \sigma(x,a,x)(a) \cdots \sigma((x,a)^{t-1}, x)(a)$. (iii) For $k = 1, \ldots, t-1$, we have $\mathbb{P}_{x\sigma}(a^k b) = \sigma(x)(a) \cdots \sigma((x,a)^{k-1}, x)(a) \cdot \sigma((x,a)^k, x)(b)$. Similarly, the strategy σ' also induces a probability distribution on W^t. It follows on the finite horizon t that

$$\begin{aligned}\mathbb{P}_{x\sigma}(t^* \leq t) &= \sum_{\tau \in W^t} \mathbb{P}_{x\sigma}(\tau) \cdot \mathbb{P}_{x\tau}(t^* \leq t) \\ &= \mathbb{P}_{x\sigma}(a^{t-2}b) \cdot \left(1 - \frac{1}{2^t}\right) + (1 - \mathbb{P}_{x\sigma}(a^{t-2}b)) \cdot \left(1 - \frac{1}{2^{t-1}}\right),\end{aligned}$$

and similarly for the strategy σ'.

Thus, $\mathbb{P}_{x\sigma}(t^* \leq t) > \mathbb{P}_{x\sigma'}(t^* \leq t)$ if and only if $\mathbb{P}_{x\sigma}(a^{t-2}b) > \mathbb{P}_{x\sigma'}(a^{t-2}b)$, if and only if $\mathbb{P}_{x\sigma}(a_{t-1} = b) > \mathbb{P}_{x\sigma'}(a_{t-1} = b)$. The proof is complete. □

By Claim 2, the stationary strategy $(\frac{1}{2}, \frac{1}{2})^\infty$ that always chooses action a and action b each with probability $\frac{1}{2}$ overtakes every pure strategy. Also, the stationary strategy $(p, 1-p)^\infty$ overtakes the stationary strategy $(q, 1-q)^\infty$ if $q < p < 1$, as for large periods t we have

$$\begin{aligned}\mathbb{P}_{x,(p,1-p)^\infty}(a_{t-1} = b) &= p^{t-2} \cdot \left(\frac{1}{2}\right)^{t-2} \cdot (1-p) > q^{t-2} \cdot \left(\frac{1}{2}\right)^{t-2} \cdot (1-q) \\ &= \mathbb{P}_{x,(q,1-q)^\infty}(a_{t-1} = b).\end{aligned}$$

This means that in Example 4.1 there is no stationary overtaking optimal strategy. We now show that there is no overtaking optimal strategy at all.

Claim 3 The MDP in Example 4.1 admits no overtaking optimal strategy.

Proof Consider any strategy σ. We construct a strategy that overtakes σ. The strategy σ can be seen as a sequence $(\xi_n)_{n=1}^{\infty}$ where ξ_n denotes the probability that σ assigns to action b when being in state x at period n. We distinguish two cases.

Case 1 Assume that either $\xi_n = 0$ for all periods n or $\xi_n = 1$ for some period n. In this case, $\mathbb{P}_{x\sigma}(a_n = b) = 0$ at large periods n. Hence, by Claim 2, the stationary strategy $(\frac{1}{2}, \frac{1}{2})^{\infty}$ overtakes σ.
Case 2 Assume that $\xi_m > 0$ for some period m and $\xi_n < 1$ for all periods n. We can choose a sequence $(\xi'_n)_{n=1}^{\infty}$ such that (i) for all periods $n = 1, \ldots, m-1$ we have $\xi'_n = \xi_n$, (ii) for period m we have $\xi'_m < \xi_m$, (iii) for all periods $n > m$ we have $\xi_n < \xi'_n < 1$, and (iv) $\prod_{n=1}^{\infty}(1-\xi'_n) = \prod_{n=1}^{\infty}(1-\xi_n)$. The idea is to slightly reduce the probability ξ_m at period m and slightly increase all probabilities ξ_n, $n > m$, so that (iv) holds, i.e., the total probability of ever playing b under $(\xi_n)_{n=1}^{\infty}$ is equal to that under $(\xi'_n)_{n=1}^{\infty}$.

Let σ' be the strategy corresponding to $(\xi'_n)_{n=1}^{\infty}$. Consider a period $t > m$. By (iii), we have $\prod_{n=t}^{\infty}(1-\xi'_n) \leq \prod_{n=t}^{\infty}(1-\xi_n)$. Hence, by (iv), we obtain $\prod_{n=1}^{t-1}(1-\xi'_n) \geq \prod_{n=1}^{t-1}(1-\xi_n)$. This means that the probability of being in state x at period t is at least as large under σ' as under σ. Thus, by (iii), we obtain $\mathbb{P}_{x\sigma'}(a_t = b) > \mathbb{P}_{x\sigma}(a_t = b)$. Since this is true for all periods $t > m$, in view of Claim 2, σ' overtakes σ. □

The following example, which is an adaptation of Example 4.1, shows that if the MDP with reachability objective is not piecewise deterministic, then it can happen that each pure strategy is overtaken by another pure strategy. As a consequence, Theorem 4.1 cannot be extended to all MDPs.

Example 4.2 Consider the MDP with initial state x and reachability objective that is depicted in Fig.3. In this MDP, the only choice of the decision maker is when to play action c, if at all, and action c can be played at most once.
In this MDP, action c leads to the target state s^* with probability $5/8$, to state y with probability $1/8$ and to state x' with probability $1/4$. It will be easier to think about action c in the following way, which gives the same transition probabilities: After playing action c, a lottery is executed: (1) with probability $1/2$ the play follows the upper-part of the arrow, and thus the play moves to state s^* with probability $3/4$ and to state y with probability $1/8$, and (2) with probability $1/2$ the play follows the bottom-part of the arrow, and thus the play moves to state s^* with probability $1/2$ and to state x' with probability $1/2$. Action c' in state x' has a similar interpretation. Note that this MDP is not piecewise deterministic, as each of the actions c and c' leads to two non-target states with a positive probability.

The pure strategies in this MDP are $a^{\infty}, c, ac, a^2c, \ldots$. The strategy a^{∞} corresponds to strategy a^{∞} in Example 4.1, and the strategy $a^t c$ corresponds to the mixed strategy in Example 4.1 that, in state x, recommends action a up to period t and the mixed action $(\frac{1}{2}, \frac{1}{2})$ at all periods after t. The reader can verify that a^{∞} is overtaken by c, and each $a^t c$ is overtaken by $a^{t+1}c$. That is, each pure strategy is overtaken by another pure strategy.

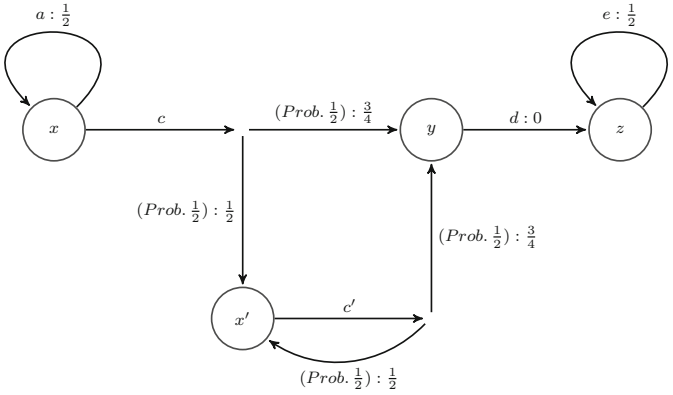

Fig. 3 The MDP in Example 4.2

5 Reachability Objectives: Generic MDPs

As shown in Sect. 4, an overtaking optimal strategy may fail to exist. In this section, we show that this is due to non-genericity of the transition function: when transitions are generic, an overtaking optimal strategy always exists. Generic transitions are natural in various applications, where transitions are affected by random noise.

We call an MDP with reachability objective *generic* if (a) all transitions from states that are not the target state [3] are positive: $p(z \mid s, a) > 0$ for every $s \neq s^*$, $a \in A(s)$, and $z \in S$, and (b) the second largest eigenvalue of $A_1, A_2, \ldots, A_K$ are all different, where K is the number of pure stationary strategies, and A_j is the transition matrix of the Markov chain on the state space S induces by the j'th pure stationary strategy, for every $j \in \{1, 2, \ldots, K\}$. These requirements involve only finitely many linear equalities and inequalities. It follows that if one randomly chooses the transition function of the MDP from $(\Delta(S))^{\sum_{s \neq s^*} |A(s)|}$ according to some probability distribution that is absolutely continuous w.r.t. the Lebesgue measure, then with probability 1 the transition function is generic.

The main result of this section is the following theorem.

Theorem 5.1 *In generic MDPs with reachability objective, there is a pure stationary strategy that overtakes each other stationary strategy at each initial state.*

The idea of the proof of Theorem 5.1 is as follows. Fix a state space $S = \{1, \ldots, n\}$, where $n \geq 2$, and let the target state be state $s^* = n$. As above, we assume without loss of generality that state n is absorbing. Every stationary strategy σ defines a transition matrix A_σ. Under the strategy σ, the rate of absorption to state n is exactly $\lambda_2(A_\sigma)$, the second largest eigenvalue of A_σ. Thus, if $\lambda_2(A_\sigma) < \lambda_2(A_{\sigma'})$, then σ overtakes σ' for the reachability objective. As before, let $A_1, A_2, \ldots, A_K$ be all transition matrices that are induced by pure stationary strategies. Since the MDP is generic, the second largest eigenvalues of these matrices differ, and therefore there is one of them, say, A_1, whose second largest eigenvalue is minimal. The matrix A_σ is in the convex hull of the

[3] We can assume w.l.o.g. that state s^* is absorbing.

matrices $A_1, A_2, \ldots, A_K$, and we will prove that if $A_\sigma \neq A_1$ then $\lambda_2(A_\sigma) > \lambda_2(A_1)$. This will imply that the pure stationary strategy that corresponds to the matrix A_1 overtakes each other stationary strategy at each initial state. The proof of Theorem 5.1 consists of four steps.

Step 1 Proving that the second largest eigenvalue determines the overtaking relation between stationary strategies: When comparing two stationary strategies σ_A and σ_B generating the respective transition matrices A and B, $\lambda_2(A) < \lambda_2(B)$ implies that σ_A overtakes σ_B at each initial state.
Step 2 Proving that the second largest eigenvalue of a transition matrix A, corresponding to a stationary strategy, is equal to the largest eigenvalue of the $(n-1) \times (n-1)$ submatrix A' that remains when we remove from A the column and the row associated with the target state: $\lambda_2(A) = \lambda_1(A')$.
Step 3 Proving that for two positive square matrices A and B that differ only in one row, the largest eigenvalue of their any convex combination cannot be lower than the minimum between the largest eigenvalues of the two matrices: for every $\alpha \in]0, 1[$ we have $\lambda_1(\alpha A + (1-\alpha)B) \geq \min\{\lambda_1(A), \lambda_1(B)\}$, and if $\lambda(A) \neq \lambda(B)$ then the inequality is strict.
Step 4 Proving that it suffices to consider only matrices that differ in one row.

Steps 1–4 imply that the pure stationary strategy that corresponds to the transition matrix with minimal second largest eigenvalue overtakes each other stationary strategy, at each initial state.

Proof of Step 1: This is a slightly stronger version of Theorem 3 in [21]. [4] □

Proof of Step 2: Let σ be a stationary strategy with an $n \times n$ transition matrix A, with entry (i, j) being the probability under σ of moving from state i to state j. Since the target state $s^* = n$ is absorbing and since the sum of entries in each row is equal to 1, the largest eigenvalue of A is $\lambda_1(A) = 1$ with eigenvector $(0, 0, \ldots, 0, 1)$.

Consider the submatrix A' of A that arises when we delete the last column and the last row (which correspond to the target state). Since the MDP is generic, the matrix A' is positive, hence by the Perron–Frobenius Theorem, the largest eigenvalue $\lambda_1(A')$ of the matrix A' is a real number. As the sum of entries in each row of A' is strictly less than 1, we have $\lambda_1(A') < 1$.

The proof that $\lambda_1(A')$ is the second largest eigenvalue of the matrix A follows from the following two observations:

(i) Any eigenvalue of A' is also an eigenvalue of A. Indeed, let μ be an eigenvalue of A' with right eigenvector $(y_1, \ldots, y_{n-1})$.[5] Then μ is an eigenvalue of A with right eigenvector $(y_1, \ldots, y_{n-1}, 0)$.
(ii) If $\mu \neq 1$ is an eigenvalue of A, then μ is also an eigenvalue of A'. Indeed, let $y = (y_1, \ldots, y_n)$ be a right eigenvector of A corresponding to μ. Then, $Ay = \mu y$.

[4] This theorem in [21] implies for every initial state $s = 1, \ldots, n-1$ that if $t \in \mathbb{N}$ is large then $\mathbb{P}_{s,\sigma_A}(t^* \leq t) \geq \mathbb{P}_{s,\sigma_B}(t^* \leq t)$. However, their proof can be easily adapted to show that $\mathbb{P}_{s,\sigma_A}(t^* \leq t) > \mathbb{P}_{s,\sigma_B}(t^* \leq t)$ for large $t \in \mathbb{N}$, so that strategy σ_A overtakes strategy σ_B. Indeed, the inequalities (A.1) and (A.2) do not only imply inequality (A.3), but they actually imply a strict inequality.

[5] Recall that the same set of eigenvalues correspond both to right and left eigenvectors.

This implies $y_n = \mu \cdot y_n$, which is only possible if $y_n = 0$. Hence, μ is an eigenvalue of A' with eigenvector $(y_1, \ldots, y_{n-1})$.

□

Proof of Step 3: The statement of Step 3 follows from the following theorem. □

Theorem 5.2 *Let A and B be two positive (all elements are positive) square matrices of the same size that differ only in the first row.*[6] *For every $\alpha \in]0, 1[$ define $M_\alpha := \alpha A + (1-\alpha)B$. Then, $\lambda_1(M_\alpha) \geq min\{\lambda_1(A), \lambda_1(B)\}$, and, if $\lambda_1(A) \neq \lambda_1(B)$, then $\lambda_1(M_\alpha) > min\{\lambda_1(A), \lambda_1(B)\}$.*

Proof of Step 4: Let $A_1, \ldots, A_K$ be all transition matrices that are induced by pure stationary strategies. Since the MDP is generic, the second largest eigenvalues of these matrices are all different. Assume $\lambda_2(A_1) < \lambda_2(A_i)$ for all $i = 2, \ldots, K$. Let σ be the pure stationary strategy corresponding to A_1.

Let τ be a stationary strategy, and let A be the transition matrix corresponding to τ. For every $r = 0, 1, \ldots, n-1$ let $\mathcal{B}_r$ be the collection of all matrices that coincide with A in the first r rows and coincide with one of the matrices $A_1, A_2, \ldots, A_K$ in the other $n - r$ rows. Note that $\mathcal{B}_0 = \{A_1, A_2, \ldots, A_K\}$. Using Step 2 together with Step 3 inductively, we obtain that

$$\lambda_2(A) \geq \min_{B \in \mathcal{B}_{n-1}} \lambda_2(B) \geq \min_{B \in \mathcal{B}_{n-2}} \lambda_2(B) \geq \cdots \geq \min_{B \in \mathcal{B}_0} \lambda_2(B). \tag{2}$$

Moreover, if $A \notin \mathcal{B}_0$, then at least one of the inequalities in Eq. (2) is strict.

Now assume that $\tau \neq \sigma$. If $A \notin \mathcal{B}_0$ then $\lambda_2(A) > \min_{B \in \mathcal{B}_0} \lambda_2(B) = \lambda_2(A_1)$, whereas if $A \in \mathcal{B}_0$ then $\lambda_2(A) > \lambda_2(A_1)$ by the choice of A_1. Thus, by Step 1, σ overtakes τ at each initial state. □

Remark 5.1 Let σ be a strategy as in Theorem 5.1, $\sigma' \neq \sigma$ be a stationary strategy and s be the initial state. One can compute a horizon T such that σ outperforms σ' beyond T, i.e., $\mathbb{P}_{s\sigma}(t^* \leq t) > \mathbb{P}_{s\sigma'}(t^* \leq t)$ for all $t \geq T$, by using inequalities (A1) and (A2) in [21].

6 Sufficient Conditions for Strong Overtaking Optimality

In a discounted MDP, if for every $s \in S$ the strategy σ_s is an optimal strategy at the initial state s, then the stationary strategy that plays at each state s the mixed action that σ_s plays at the initial period is also optimal. The next theorem aims at developing the analogous result for strongly overtaking optimal strategies in MDPs with reachability objective.

Theorem 6.1 *Consider an MDP with reachability objective. Suppose that, for every initial state $s \in S$, there is a strategy σ_s with the following properties:*

[6] Note that if A and B differ in several rows, then the statement is no longer true: for the $3 \times 3-$matrices $A = ((98, 98, 1), (98, 1, 1), (1, 1, 1))$ and $B = ((1, 1, 1), (1, 1, 98), (1, 98, 98))$ we have $\lambda_1(A) = \lambda_1(B) \approx 158.86$, while $\lambda_1\left(\frac{1}{2}A + \frac{1}{2}B\right) = 100$.

(i) *The strategy σ_s is strongly overtaking optimal at the initial state s.*
(ii) *Denote by α_s the mixed action that σ_s uses at period 1 in state s. The strategy σ_s weakly overtakes each strategy that uses a mixed action different from α_s at period 1 in the initial state s.*

Let α be the stationary strategy that uses the mixed action α_s at state s, for all $s \in S$. Then, α is strongly overtaking optimal at each initial state.

Proof We will use a dynamic programming argument. Fix an initial state $s \in S$. For each state $z \in S$, let $H_s(z)$ denote the set of histories h such that (1) h has a positive probability under σ_s, and (2) h ends in state z.

We show that for each $h \in H_s(z)$ we have $\sigma_s(h) = \alpha_z$. Let $h \in H_s(z)$. Suppose by way of contradiction that $\sigma_s(h) \neq \alpha_z$. Let σ_s' be the strategy such that (i) σ_s' follows σ_s outside the subgame that starts at h, and (ii) in the subgame that starts at h, the continuation strategy $\sigma[h]$ is replaced by σ_z. Then, for each period t that is larger than the last period in the history h we have $\mathbb{P}_{\sigma_s'}(t^* \leq t) - \mathbb{P}_{\sigma_s}(t^* \leq t) = \mathbb{P}_{\sigma_s}(h) \cdot \left[\mathbb{P}_{z,\sigma_z}(t^* \leq t) - \mathbb{P}_{z,\sigma[h]}(t^* \leq t)\right]$. By (ii), σ_z weakly overtakes $\sigma[h]$ for initial state z. Therefore, the quantity $\mathbb{P}_{z,\sigma_z}(t^* \leq t) - \mathbb{P}_{z,\sigma[h]}(t^* \leq t)$ is non-negative for all large t and strictly positive for infinitely many t. Thus, the same holds for $\mathbb{P}_{\sigma_s'}(t^* \leq t) - \mathbb{P}_{\sigma_s}(t^* \leq t)$, and hence σ_s is weakly overtaken by σ_s'. This is a contradiction to (i).

Hence, each history h has the same probability under σ_s and under α. As σ_s is strongly overtaking optimal at the initial state s, so is the strategy α. □

7 Safety Objectives

The model of MDPs with safety objective is similar to the model of MDPs with reachability objective, except that the decision maker's objective is to reach the state s^* with as low a probability as possible.

Overtaking optimality. A strategy σ *overtakes* a strategy σ' at the initial state s if there is $T \in \mathbb{N}$ such that $\mathbb{P}_{s\sigma}(t^* \leq t) < \mathbb{P}_{s\sigma'}(t^* \leq t)$ for all $t \geq T$. A strategy σ is *overtaking optimal* at the initial state s if there is no strategy that overtakes σ at that initial state.

Strong overtaking optimality. A strategy σ *weakly overtakes* a strategy σ' at the initial state s if there is $T \in \mathbb{N}$ such that for all $t \geq T$ we have $\mathbb{P}_{s\sigma}(t^* \leq t) \leq \mathbb{P}_{s\sigma'}(t^* \leq t)$ with strict inequality for infinitely many t. If σ overtakes σ' at the initial state s then σ also weakly overtakes σ' at that initial state. A strategy σ is *strongly overtaking optimal* at the initial state s if no strategy weakly overtakes σ at that initial state. A strongly overtaking optimal strategy at the initial state s is also overtaking optimal at that state.

Results. Theorem 4.1 remains valid for safety objectives. The proof requires the following changes. (1) We can still assume that the MDP $\mathcal{M}$ has no state $s \neq s^*$ and action $a \in A(s)$ with $p(s^* \mid s, a) = 1$. Indeed, such an action can be deleted, and if all actions in a state s are deleted, then we can delete the state s and replace each

transition to s by a transition to s^*. (2) Because now the decision maker prefers low probabilities to state s^*, the payoffs in the auxiliary MDP $\mathcal{M}'$ are defined to be the opposite: $u'(s,a) = \log(p(w(s,a) \mid s,a)),$.

Theorems 5.1 and 6.1 remain valid for safety objectives, with analogous proofs. Similarly to Example 4.1, the following MDP with safety objective has no overtaking optimal strategy: Take the MDP in Example 4.1 and replace $b : \frac{3}{4}$ with $b : 0$ and $d : 0$ with $d : \frac{3}{4}$. In this MDP, b is still preferred over d.

8 Conclusions

It remains an open problem if generic MDPs with reachability objective admit a pure stationary strategy that is strongly overtaking optimal (cf. Theorem 5.1). The difficulty is that when we allow non-stationary strategies, the transition probabilities generally cannot be described by a single transition matrix.

When using a stationary strategy, sometimes it is important to study the probability distribution of the current state, at any period t, on condition that the state s^* has not been reached yet. This conditional distribution converges under some conditions to a limit, called a quasi-stationary distribution. This convergence and its speed are subject of study in the literature; see, e.g., [22].

Acknowledgements We are grateful to Laurent Miclo for a helpful discussion and for referring us to the related literature in linear algebra, to Mickael Randour for a helpful discussion and his comments on a draft of this paper, to Robert Israel for providing the proof of Theorem 5.2, to Emilio De Santis and Fabio Spizzichino for providing us with the right references, and to Antonín Kučera for a valuable discussion. Finally, we thank two anonymous referees for the careful reading of our paper and their comments. This work has been partly supported by COST Action CA16228 European Network for Game Theory. Ashkenazi-Golan and Solan acknowledge the support of the Israel Science Foundation, Grants #217/17 and #722/18, and the NSFC-ISF Grant #2510/17.

Appendix: The proof of Theorem 4.2.

We closely follow the proof of Theorem 1 (that [1] implies [4], page 219) in [6]. Fix the initial state s.

Since we can add a constant to all payoffs, we can assume that $v(s) = 0$.

A sequence $\ell = (s_1, a_1, s_2, \ldots, s_t, a_t, s_{t+1})$ is called a *loop* if (1) $s_1, \ldots, s_t$ are distinct elements of S, (2) $s_1 = s_{t+1}$, (3) $a_i \in A(s_i)$ for all $i = 1, \ldots, t$, and (4) $p(s_{i+1} \mid s_i, a_i) = 1$ for all $i = 1, \ldots, t$. For a loop $\ell = (s_1, a_1, \ldots, s_t, a_t, s_{t+1})$ let $\phi(\ell)$ denote the sum of the payoffs along ℓ, i.e., $\phi(\ell) = \sum_{n=1}^{t} u(s_t, a_t)$. The number

of loops is finite and therefore the following quantity is negative: $\delta = \max\{\phi(\ell) : \ell \text{ is a loop and } \phi(\ell) < 0\}$; the definition of δ is irrelevant if the set over which the maximum is taken is empty.

Let σ be a Blackwell optimal strategy for the initial state s, and σ' be a pure strategy. Let $s_1 = s, a_1, s_2, a_2, \ldots$ and $s'_1 = s, a'_1, s'_2, a'_2, \ldots$ be the sequences of states and actions induced by σ and respectively by σ'. Denote $u_t = u(s_t, a_t)$ and $u'_t = u(s'_t, a'_t)$. The following two claims are proven in [6], page 220.

CLAIM 1: Suppose that $s_k = s_{k+m+1}$ for some $k, m \in \mathbb{N}$. Then $u_k + \cdots + u_{k+m} = 0$.
CLAIM 2: Suppose that $s'_k = s'_{k+m+1}$ for some $k, m \in \mathbb{N}$. Then either $u'_k + \cdots + u'_{k+m} = 0$ or $u'_k + \cdots + u'_{k+m} \leq \delta$ (recall that $\delta < 0$).

Suppose by way of contradiction that there is $M \in \mathbb{N}$ such that $u_T(s, \sigma') \geq u_T(s, \sigma)$ for all $T \geq M$ and the inequality is strict for an infinite sequence $T_1 < T_2 < \ldots$, where $M < T_1$. Let $q_1 = 0$ and $q_t = u_1 + \cdots + u_{t-1}$ for each $t \geq 1$. Likewise, let $q'_1 = 0$ and $q'_t = u'_1 + \cdots + u'_{t-1}$ for each $t \geq 1$. Due to our assumption about σ', we have $q'_t - q_t > 0$ for each $t = T_1, T_2, \ldots$.

We next argue that there exists $\mu > 0$ such that $q'_t - q_t \geq \mu$ for each $t = T_1, T_2, \ldots$. Indeed, suppose that no such $\mu > 0$ exists. Then, by taking a subsequence if necessary, we can assume that the sequence $\{q'_{T_n} - q_{T_n}\}_{n\in\mathbb{N}}$ is strictly decreasing: $q'_{T_n} - q_{T_n} > q'_{T_{n+1}} - q_{T_{n+1}}$ for every $n \in \mathbb{N}$. As is shown in [6], page 220, by claims 1 and 2 above, this leads to $q'_{T_n} - q_{T_n} < 0$ for all sufficiently large n, which is a contradiction.

Denote by $u_\beta(s, \sigma)$ and $u_\beta(s, \sigma')$ the expected β-discounted payoffs under σ and σ'. So,

$$\begin{aligned}\frac{u_\beta(s,\sigma)}{1-\beta} &= \sum_{t=1}^{\infty} \beta^{t-1} u_t = \sum_{t=1}^{\infty} \beta^{t-1}(q_{t+1} - q_t) \\ &= \sum_{t=2}^{\infty} \beta^{t-2} q_t - \sum_{t=1}^{\infty} \beta^{t-1} q_t = \sum_{t=2}^{\infty} (\beta^{t-2} - \beta^{t-1}) q_t.\end{aligned}$$

Likewise, $\frac{u_\beta(s,\sigma')}{1-\beta} = \sum_{t=2}^{\infty}(\beta^{t-2} - \beta^{t-1})q'_t$. Take any $\beta \in]0, 1[$ for which σ is β-optimal for the initial state s. We use the notation $D^* := 2 \cdot \max_{s\in S, a\in A(s)} |u(s, a)|$. Then, $q'_t - q_t \geq -(t-1)D^*$, for every $t \geq 1$. Note that by Property I, $q'_t - q_t \geq 0$ for every $t \geq M$. Hence, we obtain

$$\begin{aligned}0 \geq \frac{u_\beta(s,\sigma') - u_\beta(s,\sigma)}{1-\beta} &= \sum_{t=2}^{\infty} (\beta^{t-2} - \beta^{t-1})(q'_t - q_t) \\ = \sum_{t=2}^{T_1} (\beta^{t-2} - \beta^{t-1})(q'_t - q_t) &+ \sum_{k=1}^{\infty} \sum_{t=T_k+1}^{T_{k+1}} (\beta^{t-2} - \beta^{t-1})(q'_t - q_t) \\ \geq -(1-\beta) \sum_{t=2}^{T_1} \beta^{t-2} \cdot t \cdot D^* &+ \sum_{k=1}^{\infty} \sum_{t=T_k+1}^{T_{k+1}-1} (\beta^{t-2} - \beta^{t-1})(q'_t - q_t)\end{aligned}$$

$$+\sum_{k=1}^{\infty}(\beta^{T_k-2}-\beta^{T_k-1})(q'_{T_k}-q_{T_k})$$
$$>(1-\beta)\left(-(T_1)^2D^*+\sum_{k=1}^{\infty}\beta^{T_k-2}\mu\right).$$

We deduce that $\sum_{k=1}^{\infty}\beta^{T_k-2}\mu < (T_1)^2D^*$ for every β sufficiently close to 1, which is a contradiction, since $\lim_{\beta\uparrow 1}\sum_{k=1}^{\infty}\beta^{T_k-2}\mu=\infty$. □

The proof of Theorem 5.2 Denote $\lambda^* = \lambda_1(M_\alpha)$. By the Perron–Frobenius theorem, there exists a positive right eigenvector $u=(u_j)_j$ corresponding to λ^* for the matrix M_α. Let e_j be the j'th unit vector. Note that u_j is the j'th coordinate of u, while e_j is the j'th unit vector. For every $j\neq 1$ we have $e_j^TA=e_j^TB$, hence $e_j^TAu=e_j^TBu$. Since $\lambda^*u_j=\lambda^*e_j^Tu=e_j^TM_\alpha u=\alpha e_j^TAu+(1-\alpha)e_j^TBu$, we deduce that $e_j^TAu=e_j^TBu=\lambda^*u_j$ for every $j\neq 1$. For $j=1$, we have $\lambda^*u_1=\lambda^*e_1^Tu=e_1^TM_\alpha u=\alpha e_1^TAu+(1-\alpha)e_1^TBu$. Assume without loss of generality that $e_1^TAu\geq e_1^TBu$, so that $e_1^TAu\geq\lambda^*u_1\geq e_1^TBu$. It follows that for every nonnegative vector v we have

$$v^TAu\geq\lambda^*v^Tu\geq v^TBu. \tag{3}$$

By the Perron–Frobenius theorem, there exists a positive left eigenvector v_A for $\lambda_1(A)$, and a positive left eigenvector v_B for $\lambda_1(B)$. Substituting $v=v_A$ in the left-hand side of Ineq. (3) and substituting $v=v_B$ in the right-hand side of Ineq. (3) we obtain

$$\lambda_1(A)v_A^Tu=v_A^TAu\geq\lambda^*v_A^Tu\quad\text{and}\quad\lambda_1(B)v_B^Tu=v_B^TBu\leq\lambda^*v_B^Tu. \tag{4}$$

Since v_A^Tu and v_B^Tu are positive reals, this implies that $\lambda_1(A)\geq\lambda^*\geq\lambda_1(B)$.

Suppose now that $\lambda_1(A)>\lambda_1(B)$. Then necessarily $e_1^TAu>e_1^TBu$, since, if the two are equal, then there would have been equality in Eq. (3), and then we would obtain that $\lambda_1(A)=\lambda_1(B)$. As all coordinates of v_A and v_B are positive, and in particular the first coordinates are positive, there is a strict inequality in both inequalities in Eq. (4). □

References

1. Baier, C., Katoen, J.P.: Principles of Model Checking. MIT Press, Cambridge (2008)
2. Chatterjee, K., Henzinger, T.A.: A survey of stochastic ω-regular games. J. Comput. Syst. Sci. **78**, 394–413 (2012)
3. Bruyère, V.: Computer aided synthesis: a game-theoretic approach. In: Charlier, E., Leroy, J., Rigo, M. (eds.) Developments in Language Theory. Lecture Notes in Computer Science, vol. 10396, pp. 3–35. Springer, Berlin (2017)
4. Howard, R.A.: Dynamic Programming and Markov Processes. MIT Press, Cambridge (1960)
5. Blackwell, D.: Discrete dynamic programming. Ann. Math. Stat. **33**, 719–726 (1962)
6. Flesch, J., Predtetchinski, A., Solan, E.: Sporadic overtaking optimality in Markov decision problems. Dyn. Games Appl. **7**, 212–228 (2017)

7. Randour, M., Raskin, J.-F., Sankur, O.: Variations on the stochastic shortest path problem. In: International Workshop on Verification, Model Checking, and Abstract Interpretation, pp. 1–18. Springer, Berlin, Heidelberg (2015)
8. Randour, M., Raskin, J.-F., Sankur, O.: Percentile queries in multi-dimensional Markov decision processes. Form. Methods Syst. Des. **50**(2–3), 207–248 (2017)
9. Brihaye, T., Bruyère, V., De Pril, J.: On equilibria in quantitative games with reachability/safety objectives. Theory Comput. Syst. **54**, 150–189 (2014)
10. Stern, L.E.: Criteria of optimality in the infinite-time optimal control problem. J. Optim. Theory Appl. **44**, 497–508 (1984)
11. Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley, New York (1994)
12. Carlson, D.A., Haurie, A., Leizarowitz, A.: Infinite Horizon Optimal Control. Springer-Verlag, Berlin (1991)
13. Zaslavski, A.J.: Turnpike Properties in the Calculus of Variations and Optimal Control. Springer, New York (2006)
14. Zaslavski, A.J.: Turnpike Phenomenon and Infinite Horizon Optimal Control. Springer, New York (2014)
15. Guo, X., Hernández-Lerma, O.: Continuous-Time Markov Decision Processes. Springer, Berlin (2009)
16. Méder, Z., Flesch, J., Peeters, R.: Optimal choice for finite and infinite horizons. Oper. Res. Lett. **40**, 469–474 (2012)
17. Nowak, A.S., Vega-Amaya, O.: A counterexample on overtaking optimality. Math. Method Oper. Res. **49**, 435–439 (1999)
18. Leizarowitz, A.: Overtaking and almost-sure optimality for infinite horizon Markov decision processes. Math. Oper. Res. **21**, 158–181 (1996)
19. Shapley, L.S.: Stochastic games. Proc. Natl. Acad. Sci. **39**, 1095–1100 (1953)
20. Davis, M.H.A.: Piecewise-deterministic Markov processes: a general class of non-diffusion stochastic models. J. R. Stat. Soc. Ser. B Methodol. **46**, 353–388 (1984)
21. De Santis, E., Spizzichino, F.: Usual and stochastic tail orders between hitting times for two Markov chains. Appl. Stoch. Models Bus. Ind. **32**, 526–538 (2016)
22. Diaconis, P., Laurent, M.: On quantitative convergence to quasi-stationarity. Ann. Faculté Sci. Toulouse: Math. **24**, 973–1016 (2015)