Dyn Games Appl (2017) 7:212–228
DOI 10.1007/s13235-016-0186-2

Sporadic Overtaking Optimality in Markov Decision Problems

János Flesch[1] · **Arkadi Predtetchinski**[2] · **Eilon Solan**[3]

Published online: 13 April 2016

Abstract This paper examines a notion of sporadic overtaking optimality in the context of Markov decision problems (MDP). For the class of deterministic MDPs, we prove the existence of pure stationary sporadic overtaking optimal strategies under both the discounted and the average payoff evaluations. Moreover, we examine logical connections between sporadic overtaking optimality and Blackwell optimality. In the class of nondeterministic MDPs, we give examples that admit no sporadic overtaking optimal strategy and discuss a number of alternative definitions of this concept.

Keywords Overtaking optimality · Sporadic overtaking optimality · Markov decision problems · Blackwell optimality

1 Introduction

In the literature on dynamic optimization problems with infinite horizon, a number of evaluation criteria have been examined. The most common ones are the evaluations by means

We would like to thank Zsombor Méder and Ronald Peeters for helpful discussions. Solan acknowledges the support of the ISF Grant #323/13.

✉ Arkadi Predtetchinski
a.predtetchinski@maastrichtuniversity.nl

János Flesch
j.flesch@maastrichtuniversity.nl

Eilon Solan
eilons@post.tau.ac.il

[1] Department of Quantitative Economics, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands

[2] Department of Economics, Maastricht University, P.O. Box 616, 6200 MD Maastricht, The Netherlands

[3] School of Mathematical Sciences, Tel-Aviv University, 69978 Tel Aviv, Israel

of the discounted sum of payoffs and the long-term average payoffs. Under either type of evaluation, optimal strategies are generically not unique. Moreover, optimality of the strategy on the infinite horizon does not guarantee its optimality for finite horizons. In fact, it is possible that a strategy that is optimal over the infinite horizon is outperformed by another optimal strategy on every finite horizon. To take this point into account, various definitions of overtaking optimality have been proposed. For an overview of some of these concepts, we refer to [2,3,6,8,10–12].

In this paper, we focus on sporadic overtaking optimality as defined in Stern [10] in the context of Markov decision problems. We only consider pure strategies. A strategy is sporadic overtaking (SO) optimal if there is no other strategy that outperforms it on all sufficiently long but finite horizons. SO optimality is a refinement of optimality: An SO optimal strategy is always optimal on the infinite horizon. The converse need not hold.

One important feature of the definition of SO optimality is that it does not require the strategy to outperform all other strategies on finite horizons. It only requires that the strategy is not outperformed on all long finite horizons by any other strategy. SO optimality is therefore weaker than overtaking optimality and uniform overtaking optimality as in Stern [10] and weaker than strong overtaking optimality as in Nowak and Vega-Amaya [7] or in Leizarowitz [5]. See also Méder et al. [6], who delineate "not-outperformed" definitions of optimality from "outperform-all" definitions of optimality.

In this paper, we study the existence of SO optimal strategies. Méder et al. [6] conjecture that an SO optimal strategy always exists in the case of the discounted evaluation,[1]; however, they also show that an approach to establishing existence using Zorn's lemma runs into difficulties. In this paper, we prove the conjecture for the class of deterministic MDPs for both the discounted payoff and the average payoff evaluations. We construct an SO optimal strategy explicitly for the discounted payoff evaluation. The idea is to split the play of the MDP into two phases: In the first phase, the decision maker collects "good" payoffs, and in the second phase, he maintains them by going through a loop of states and actions. The strategy thus constructed is stationary and independent of the initial state. We show that, if the discount factor is large enough, the same strategy is also SO optimal for the average payoff evaluation.

In the class of nondeterministic MDPs, we provide a counterexample to the existence of SO discounted optimal strategies and a counterexample to the existence of SO average optimal strategies.

Furthermore, we examine logical connections between overtaking optimality and Blackwell optimality. We show that a strategy is Blackwell optimal if and only if it is SO discounted optimal for large discount factors. Furthermore, both of these conditions imply that the strategy is SO average optimal. The converse does not hold.

The structure of the paper. In Sect. 2, we describe the model and provide the main definitions. In Sect. 3, we state our main results. Sections 4 and 5 contain the proof of the main results. In Sect. 6, we show that the existence result does not extend to nondeterministic MDPs. In Sect. 7, we discuss two extensions of the model: In Sect. 7.1, we examine alternative definitions of SO optimality, and in Sect. 7.2, we examine SO optimality in randomized strategies.

2 The Model

Markov decision problems. A *Markov decision problem*, MDP for brevity, is given by [1] a nonempty and finite set S of states, [2] for each state $s \in S$, a nonempty and finite set $A(s)$

[1] In Méder et al. [6], SO optimal strategies are called repeatedly optimal.

of actions, [3] for each state $s \in S$ and each action $a \in A(s)$, a payoff $r(s,a) \in \mathbb{R}$, and [4] for each state $s \in S$ and each action $a \in A(s)$, a probability distribution $p(s,a)$ on the set S of states.

An MDP is played at stages in $\mathbb{N}$ as follows: At stage 1, in a given initial state s_1, the decision maker chooses an action $a_1 \in A(s_1)$. Then, the decision maker receives payoff $r(s_1,a_1)$, and subsequently state $s_2 \in S$ is drawn from the distribution $p(s_1,a_1)$. At stage 2, in state s_2, the decision maker chooses an action $a_2 \in A(s_2)$, receives payoff $r(s_2,a_2)$, and subsequently state s_3 is drawn from $p(s_2,a_2)$, and so on.

We say that an MDP is *deterministic* if for every state $s \in S$ and every available action $a \in A(s)$, there is a state $s' \in S$ such that $p(s'|s,a) = 1$. Otherwise, the MDP is called *nondeterministic*.

The *history at stage* $t \in \mathbb{N}$ is a sequence $(s_1, a_1, \ldots, s_{t-1}, a_{t-1}, s_t)$ such that $s_i \in S$ for all $i = 1, \ldots, t$, and $a_i \in A(s_i)$ and $p(s_{i+1}|s_i, a_i) > 0$ for all $i = 1, \ldots, t-1$. We denote by H_t the set of all histories at stage t and by $H = \cup_{t\in\mathbb{N}} H_t$ the set of all histories. Let $s(h)$ denote the final state of each history $h \in H$. 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 $p(s_{i+1}|s_i, a_i) > 0$ for each $i \in \mathbb{N}$.

A *pure strategy* σ is a map that, to each history $h \in H$, assigns an action $\sigma(h) \in A(s(h))$. The interpretation is that σ recommends action $\sigma(h)$ if history h arises. We will only consider pure strategies (with the exception of Sect. 7.2), and therefore, the qualification pure is henceforth omitted. Starting from the initial state s, a strategy σ induces a probability measure on H_∞, where H_∞ is endowed with the sigma-algebra generated by the cylinder sets. We denote the corresponding expectation operator by $\mathbb{E}_{s\sigma}$.

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 $s(h) = s(h')$. A stationary strategy can thus be seen as an element of $\times_{s\in S} A(s)$.

Discounted Payoff and Discounted Optimality. For discount factor $\beta \in [0,1)$, initial state $s \in S$, and strategy σ, the *(normalized) β-discounted expected payoff* is given by

$$u_\beta(s,\sigma) = (1-\beta) \sum_{t=1}^{\infty} \beta^{t-1} \mathbb{E}_{s\sigma}\left(r(s_t, a_t)\right),$$

where s_t and a_t are random variables for the state and action at stage t.

The *β-discounted value for initial state s* is defined as

$$v_\beta(s) = \sup_\sigma u_\beta(s,\sigma).$$

A strategy σ is called *β-optimal for initial state s* if $u_\beta(s,\sigma) = v_\beta(s)$. By the results of Shapley [9] and Blackwell [1], for each $\beta \in [0,1)$ there exists a stationary strategy that is β-optimal for every initial state. Moreover, there exists[2] a discount factor α such that for any initial state s, the set of β-optimal strategies is independent of $\beta \in (\alpha, 1)$. We refer to the

[2] Let $A_\beta(s)$ denote the set of β-optimal actions in state s, that is, all actions $a \in A(s)$ such that

$$v_\beta(s) = (1-\beta) r(s,a) + \beta \sum_{w\in S} p(w|s,a) \cdot v_\beta(w).$$

The strategy σ is β-optimal if and only if $\sigma(h) \in A_\beta(s(h))$ for each history h reached by the strategy σ with positive probability. It is thus sufficient to show that there is $\alpha \in [0,1)$ such that $A_\beta(s) = A_\alpha(s)$ for all $s \in S$ and all $\beta \in (\alpha, 1)$. This follows from the fact that the function $\beta \mapsto v_\beta(s)$ is semi-algebraic for every $s \in S$.

smallest such discount factor α as the *Blackwell discount factor* of the MDP and we denote it by β_*.

A strategy σ is called *Blackwell optimal* for the initial state s if it is β-optimal for the initial state s for all β sufficiently close to 1. In view of the observation above, σ is Blackwell optimal for the initial state s if and only if there exists $\beta \in (\beta_*, 1)$ such that σ is β-optimal for s.

For a discount factor $\beta \in [0, 1)$, initial state $s \in S$, and strategy σ, the (normalized) *β-discounted expected payoff up to horizon $T \in \mathbb{N}$* is given by

$$u_{\beta,T}(s,\sigma) = \frac{1}{1+\cdots+\beta^{T-1}} \sum_{t=1}^{T} \beta^{t-1} \mathbb{E}_{s\sigma}\left(r(s_t,a_t)\right).$$

Average Payoff and Average Optimality. For initial state $s \in S$ and strategy σ, the *average expected payoff* is given by

$$u(s,\sigma) = \limsup_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{s\sigma}(r(s_t,a_t)).$$

The average value for initial state s is defined as

$$v(s) = \sup_{\sigma} u(s,\sigma).$$

It is known (e.g., [8]) that for each $s \in S$

$$\lim_{\beta\uparrow 1} v_\beta(s) = v(s). \tag{1}$$

A strategy σ is called *average optimal for initial state s* if $u(s,\sigma) = v(s)$. By Howard [4] and Blackwell [1] such an optimal strategy always exist, even a stationary one that is independent of the initial state.

For an initial state $s \in S$ and a strategy σ, the *average expected payoff up to horizon $T \in \mathbb{N}$* is given by

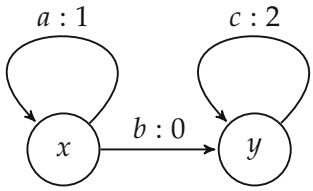
$$u_T(s,\sigma) = \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{s\sigma}(r(s_t,a_t)).$$

Sporadic Overtaking Discounted Optimality. A strategy σ is called *sporadic overtaking (SO) β-optimal for initial state s*, where $\beta \in [0, 1)$ is a discount factor, if there is no strategy σ' and finite horizon $\widetilde{T}$ such that $u_{\beta,T}(s,\sigma') > u_{\beta,T}(s,\sigma)$ for every finite horizon $T \geq \widetilde{T}$. This means that, with respect to the β-discounted payoff, there is no strategy σ' which is strictly better than σ on all long finite horizons. Equivalently, σ is SO β-optimal for initial state s, if for every strategy σ' there exists an increasing sequence of finite horizons $(T_n)_{n\in\mathbb{N}}$ such that $u_{\beta,T_n}(s,\sigma) \geq u_{\beta,T_n}(s,\sigma')$ for every $n \in \mathbb{N}$. Méder et al. [6] conjectured that every MDP admits an SO β-optimal strategy, for every initial state and every discount factor, and provided an example that points to the difficulty of this existence problem.

Note that, for any initial state s, an SO β-optimal strategy σ is also β-optimal. Indeed, take a strategy σ' that is β-optimal for the initial state s. Then, $u_{\beta,T_n}(s,\sigma) \geq u_{\beta,T_n}(s,\sigma')$ for an increasing sequence of finite horizons $(T_n)_{n\in\mathbb{N}}$. Hence, by taking the limit when n tends to infinity, we obtain $u_\beta(s,\sigma) \geq u_\beta(s,\sigma') = v_\beta(s)$.

The converse is not always true, as the following Example demonstrates.

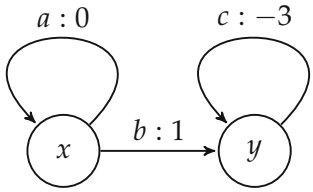
Example 1 Consider the following deterministic MDP with two states. The notation is as follows. There are two states, denoted by x and y, with x being the initial state. In state x, the decision maker has two actions. Action a gives payoff 1 and keeps the play in state x, whereas action b gives payoff 0 and brings the play to state y. In state y, the only action is c, which gives payoff 2 and keeps the play in state y.



Let $\beta = 0.5$. All strategies of the decision maker give a β-discounted payoff of 1, so all strategies are β-optimal. Yet, the only SO β-optimal strategy is the one that always chooses action a. Indeed, this strategy induces a β-discounted payoff of 1 up to any finite horizon, whereas any other strategy induces a β-discounted payoff of strictly less than 1 for all sufficiently large finite horizons.

As the example below shows, an SO β-optimal strategy need not be a maximizer of the function $u_{\beta,T}$ for any $T \in \mathbb{N}$.

Example 2 We consider a deterministic MDP with two states, similar to that in Example 1:



Here, for $\beta \in (0.5, 1)$, the only SO β-optimal strategy is to play action a all the time. This strategy yields a discounted payoff 0 up to any horizon T. However, for every $T \in \mathbb{N}$, playing a for the first $T-1$ stages and playing b at stage T yield a positive payoff up to horizon T.

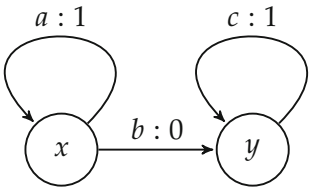
Sporadic Overtaking Average Optimality. We define sporadic overtaking (SO) average optimality analogously to the discounted case. A strategy σ is called *SO average optimal for initial state* s, if there is no strategy σ' and finite horizon $\widetilde{T}$ such that $u_T(s,\sigma') > u_T(s,\sigma)$ for every finite horizon $T \geq \widetilde{T}$. Equivalently, σ is SO average optimal for initial state s, if for every strategy σ' there exists an increasing sequence of finite horizons $(T_n)_{n\in\mathbb{N}}$ such that $u_{T_n}(s,\sigma) \geq u_{T_n}(s,\sigma')$ for every $n \in \mathbb{N}$.

Similarly to the discounted case, for any initial state s, an SO average optimal strategy σ is also average optimal. Indeed, take a stationary strategy σ' that is average optimal for the initial state s. Since σ' is stationary, $u_T(s,\sigma')$ converges to $u(s,\sigma')$ as T tends to infinity (cf., for example, [8]). By the SO average optimality of σ, we have $u_{T_n}(s,\sigma) \geq u_{T_n}(s,\sigma')$ for an increasing sequence of finite horizons $(T_n)_{n\in\mathbb{N}}$. Hence,

$$u(s,\sigma) \;\geq\; \limsup_{n\to\infty} u_{T_n}(s,\sigma) \;\geq\; \limsup_{n\to\infty} u_{T_n}(s,\sigma') \;=\; u(s,\sigma') \;=\; v(s).$$

The converse is not always true, as the following MDP demonstrates.

Example 3 Consider the following MDP with two states:



The notation is just as before, and x is again the initial state. All strategies of the decision maker give an average payoff of 1, so all strategies are average optimal. Yet, the only SO average optimal strategy is the one that always chooses action a.

Note that, similarly to the discounted case, an SO average optimal strategy does not need to be a maximizer of the function u_T, for any $T \in \mathbb{N}$.

3 Main Results

In this section, we state our main results. The first result establishes logical connections between three optimality conditions, namely Blackwell optimality, SO discounted optimality, and SO average optimality. Recall the definition of Blackwell discount factor β_* as the smallest discount factor such that for any initial state s, the set of β-optimal strategies is independent of $\beta \in (\beta_*, 1)$.

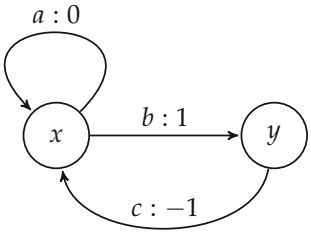
Theorem 1 *Given a deterministic Markov decision problem and a strategy σ, consider the following four statements.*

[1] *The strategy σ is Blackwell optimal for the initial state s.*
[2] *The strategy σ is SO β-optimal for the initial state s for each $\beta \in (\beta_*, 1)$.*
[3] *The strategy σ is SO β-optimal for the initial state s for some $\beta \in (\beta_*, 1)$.*
[4] *The strategy σ is SO average optimal for the initial state s.*

Statements [1]*,* [2]*, and* [3] *are equivalent, and each of them implies* [4].

We prove the above theorem in Sect. 4.
The following example shows that condition [4] does not imply [1]–[3].

Example 4 The initial state is x. Action a yields a payoff of 0 and keeps the game in state x, while action b yields a payoff of 1 and leads to state y. In state y, the only action is c, which yields the payoff -1 and leads to state x.



Always playing action a is an SO average optimal strategy. However, the only Blackwell optimal strategy and the only SO β-optimal strategy, for any $\beta \in [0, 1)$, are always playing action b in state x.

Our second result establishes the existence of SO optimal strategies in deterministic MDPs.

Theorem 2 *Consider a deterministic Markov decision problem. For each discount factor $\beta \in [0, 1)$, there exists a stationary strategy that is SO β-optimal for each initial state. Moreover, there exists a stationary strategy that is SO average optimal for each initial state.*

We prove the first part of Theorem 2 on the existence of a stationary SO discounted optimal strategy in Sect. 5. The second part of Theorem 2 follows from the implication [1] $\Rightarrow$ [4] in Theorem 1 in conjunction with Blackwell's [1] result on the existence of stationary strategies that are Blackwell optimal for all initial states. Alternatively, the second part of Theorem 2 also follows from the implication [3] $\Rightarrow$ [4] in Theorem 1 in conjunction with the first part of Theorem 2.

In Sect. 6, we argue that the above existence results cannot be extended to nondeterministic MDPs.

4 Proof of Theorem 1

Proof that [1] implies [2]. Let σ be a Blackwell optimal strategy for the initial state s. Suppose that there exists a discount factor $\beta' \in (\beta_*, 1)$ such that σ is not SO β'-optimal for s. Then there is a strategy σ' and a stage $\widetilde{T} \in \mathbb{N}$ such that $u_{\beta' T}(s, \sigma) < u_{\beta' T}(s, \sigma')$ for all $T \geq \widetilde{T}$. Taking the limit as T goes to infinity, we obtain $u_{\beta'}(s, \sigma) \leq u_{\beta'}(s, \sigma')$. Since σ is β'-optimal for the initial state s, so is σ'. In view of the definition of β_* and by the choice of β', this implies that σ' is Blackwell optimal for the initial state s. Therefore

$$u_\beta(s, \sigma) = (1 - \beta) \sum_{t=1}^{\infty} \beta^{t-1} r(s_t, a_t) = (1 - \beta) \sum_{t=1}^{\infty} \beta^{t-1} r(s'_t, a'_t) = u_\beta(s, \sigma')$$

for every $\beta \in (\beta_*, 1)$, where $s_1 = s, a_1, s_2, a_2, \ldots$ are the sequence of states and actions induced by σ, and $s'_1 = s, a'_1, s'_2, a'_2, \ldots$ are the analog sequences induced by σ'. It follows that

$$r(s_t, a_t) = r(s'_t, a'_t), \quad \forall t \in \mathbb{N}. \tag{2}$$

Indeed, the function

$$\beta \mapsto (1 - \beta) \sum_{t=1}^{\infty} \beta^{t-1} r(s_t, a_t) - (1 - \beta) \sum_{t=1}^{\infty} \beta^{t-1} r(s'_t, a'_t)$$

from the complex plain is holomorphic and has a radius of convergence at least 1. Its set of zeroes has an accumulation point at 0, and therefore, this function is zero on the unit ball. In particular, all its derivatives, which are $k!(r(s_t, a_t) - r(s'_t, a'_t))$, vanish. Equation (2) leads to a contradiction since then $u_{\beta' T}(s, \sigma) = u_{\beta' T}(s, \sigma')$ for all $T \geq T'$.

Proof that [2] implies [3]. Obvious.

Proof that [3] implies [1]. Suppose the strategy σ is SO β-optimal for the initial state s for some $\beta \in (\beta_*, 1)$. As noticed earlier, σ is then β-discounted optimal for the initial state s. It follows from the definition of β_* that σ is Blackwell optimal for the initial state s.

Proof that [1] implies [4]. Without loss of generality, 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) $p(s_{i+1}|s_i, a_i) = 1$ for all $i = 1, \ldots, t$. For a loop $\ell = (s_1, a_1, s_2, \ldots, s_t, a_t, s_{t+1})$ let $\phi(\ell)$ denote the sum of the payoffs along ℓ:

$$\phi(\ell) = \sum_{n=1}^{t} r(s_t, a_t).$$

The number of loops is finite, and therefore, the following quantity δ is negative:

$$\delta = \max\{\phi(\ell) : \ell \text{ is a loop and } \phi(\ell) < 0\} \tag{3}$$

(the definition of δ is irrelevant if the set over which the maximum is taken is empty).

Suppose that the strategy σ is Blackwell optimal for the initial state s. Let $s_1, a_1, s_2, a_2, \ldots$ be the sequence of states and actions induced by the strategy σ, where $s_1 = s$. Denote $r_t = r(s_t, a_t)$.

Suppose that there exists a strategy σ' and $M \in \mathbb{N}$ such that $u_T(s, \sigma) < u_T(s, \sigma')$ for all $T \geq M$. We let $s'_1, a'_1, s'_2, a'_2, \ldots$ be the sequence of states and actions induced by the strategy σ', where $s'_1 = s$, and denote $r'_t = r(s'_t, a'_t)$.

We first state and prove two claims, and then come back to the proof of the implication.

Claim 1. *Suppose that for some $k, m \in \mathbb{N}$ we have $s_k = s_{k+m+1}$. Then $r_k + \cdots + r_{k+m} = 0$.*

Claim 2. *Suppose that for some $k, m \in \mathbb{N}$ we have $s'_k = s'_{k+m+1}$. Then either $r'_k + \cdots + r'_{k+m} = 0$ or $r'_k + \cdots + r'_{k+m} \leq \delta$ (recall that $\delta < 0$).*

Proof of claim 1. Let $s_k = s_{k+m+1} = z$. Fix $\beta \in (\beta_*, 1)$. We know that σ is β-optimal for the initial state s and that it visits the state z at stage k. Hence, the sequence $s_k, a_k, s_{k+1}, a_{k+1}, \ldots$ that σ generates starting with stage k onward is β-optimal from the initial state z. Thus we have

$$v_\beta(s) = (1 - \beta) \sum_{t=1}^{\infty} \beta^{t-1} r_t$$

and

$$v_\beta(z) = (1 - \beta) \sum_{t=k}^{\infty} \beta^{t-k} r_t.$$

It follows that

$$v_\beta(s) = (1 - \beta) \sum_{t=1}^{k-1} \beta^{t-1} r_t + \beta^{k-1} v_\beta(z).$$

Taking the limits as $\beta \uparrow 1$ and using (1), we obtain $v(s) = v(z)$. Thus $v(z) = 0$. Similarly, since σ visits z at stage $k + m + 1$, the sequence $s_{k+m+1}, a_{k+m+1}, s_{k+m+2}, a_{k+m+2}, \ldots$ is also β-optimal for the initial state z. Thus

$$v_\beta(z) = (1 - \beta) \sum_{t=k+m+1}^{\infty} \beta^{t-k-m-1} r_t.$$

It follows that

$$v_\beta(z) = (1-\beta)\sum_{t=k}^{k+m} \beta^{t-k} r_t + \beta^{m+1} v_\beta(z).$$

Rearranging we find that

$$v_\beta(z) = \left(\frac{1-\beta}{1-\beta^{m+1}}\right)\sum_{t=k}^{k+m} \beta^{t-k} r_t.$$

Taking the limit as $\beta \uparrow 1$ and using (1), we obtain

$$v(z) = \frac{1}{m+1}(r_k + \cdots + r_{k+m}).$$

Since $v(z) = 0$, we obtain $r_k + \cdots + r_{k+m} = 0$, as desired.

Proof of claim 2. Consider a strategy σ'' that coincides with σ' up to the stage k, after which it repeats the sequence $(s'_k, a'_k, \ldots, s'_{k+m}, a'_{k+m})$ ad infinitum. One can verify that the average payoff w.r.t. σ'' for the initial state s is $u(s,\sigma'') = \frac{1}{m+1}(r'_k + \cdots + r'_{k+m})$. Since $u(s,\sigma'') \leq v(s) = 0$, we have $r'_k + \cdots + r'_{k+m} \leq 0$. Furthermore, the expression $r'_k + \cdots + r'_{k+m}$ can be written as a sum $\phi(\ell_1) + \cdots + \phi(\ell_n)$ where each ℓ_i is a loop induced by σ'. By (3), it follows that the sum $\phi(\ell_i)$ of payoffs along each loop ℓ_i is either zero or is bounded above by δ. It follows that the sum $r'_k + \cdots + r'_{k+m}$ is either 0 or is bounded above by δ.

Completion of the proof of the implication. Let $q_1 = 0$ and $q_t = r_1 + \cdots + r_{t-1}$ for each $t \geq 1$. Likewise, let $q'_1 = 0$ and $q'_t = r'_1 + \cdots + r'_{t-1}$ for each $t \geq 1$. Due to our assumption about σ', we have $q'_t - q_t > 0$ for each $t \geq M+1$. We next argue that there exists $\mu > 0$ and such that

$$q'_t - q_t \geq \mu \text{ for each } t \geq M+1. \tag{4}$$

Suppose no such $\mu > 0$ exists. Then there is a sequence $M+1 = t_0 < t_1 < \cdots$ such that $\{q'_{t_n} - q_{t_n}\}_{n\in\mathbb{N}}$ is a strictly decreasing sequence, that is $q'_{t_n} - q_{t_n} > q'_{t_{n+1}} - q_{t_{n+1}}$ for every $n \in \mathbb{N}$. Replacing the sequence $t_0, t_1, \ldots$ by a subsequence if necessary, we can assume that the sequence $\{(s_{t_n}, s'_{t_n}), n \in \mathbb{N}\}$ is constant: There are $y, z \in S$ such that $(s_{t_n}, s'_{t_n}) = (y,z)$ for all $n \in \mathbb{N}$. Take any $n \in \mathbb{N}$. Since $(s_{t_n}, s'_{t_n}) = (s_{t_{n+1}}, s'_{t_{n+1}})$, we know that $r_{t_n} + \cdots + r_{t_{n+1}-1} = 0$ and that $r'_{t_n} + \cdots + r'_{t_{n+1}-1} \leq 0$. Furthermore, the latter sum is either zero or is bounded above by δ. Notice that

$$q'_{t_{n+1}} - q_{t_{n+1}} = q'_{t_n} + \sum_{i=t_n}^{t_{n+1}-1} r'_i - q_{t_n} - \sum_{i=t_n}^{t_{n+1}-1} r_i. \tag{5}$$

Using the fact $q'_{t_n} - q_{t_n} > q'_{t_{n+1}} - q_{t_{n+1}}$ and $r_{t_n} + \cdots + r_{t_{n+1}-1} = 0$, we conclude that $r'_{t_n} + \cdots + r'_{t_{n+1}-1} < 0$. Therefore $r'_{t_n} + \cdots + r'_{t_{n+1}-1} \leq \delta$. Consequently using (5), we derive $q'_{t_n} - q_{t_n} + \delta \geq q'_{t_{n+1}} - q_{t_{n+1}}$. Since this holds for every $n \in \mathbb{N}$, and since $\delta < 0$, we conclude that $q'_{t_n} - q_{t_n} < 0$ for all n large enough, leading to a contradiction.

Now,

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

Likewise,

$$\frac{u_\beta(s,\sigma')}{1-\beta} = \sum_{t=2}^{\infty}(\beta^{t-2}-\beta^{t-1})q'_t.$$

Take any $\beta \in (\beta_*, 1)$. Using the fact that σ is β-optimal for the initial state s and (4), we obtain

$$\begin{aligned}0 &\geq \frac{u_\beta(s,\sigma') - u_\beta(s,\sigma)}{1-\beta}\\ &= \sum_{t=2}^{\infty}(\beta^{t-2}-\beta^{t-1})(q'_t - q_t)\\ &\geq \sum_{t=2}^{M}(\beta^{t-2}-\beta^{t-1})(q'_t - q_t) + \sum_{t=M+1}^{\infty}(\beta^{t-2}-\beta^{t-1})\mu\\ &= \sum_{t=2}^{M}(\beta^{t-2}-\beta^{t-1})(q'_t - q_t) + \beta^{M-1}\mu\end{aligned}$$

Finally, taking the limit of the last inequality as $\beta \uparrow 1$, we obtain $0 \geq \mu$, a contradiction.

5 Proof of Theorem 2

In this section, we prove Theorem 2. Take a deterministic MDP. As argued in Sect. 3, we only need to show that, for each discount factor $\beta \in [0, 1)$, there exists a stationary strategy that is SO β-optimal for each initial state.

Let $\beta \in [0, 1)$. Our construction leads to a stationary SO β-optimal strategy with the following structure: The strategy prescribes first to follow a path from the initial state to a specifically chosen state, and afterward, the strategy prescribes to repeat a loop indefinitely. The pair of the path and the loop is chosen such that (a) the pair maximizes the discounted payoff, and (b) subject to part (a), the pair maximizes the discounted payoff along the path.

Infinite histories. Consider a state $s \in S$. Let W_s denote the set of states $w \in S$ such that there is a history that starts at s, ends at w, and only uses β-optimal actions. Further, let L_s denote the set of states $w \in W_s$ such that there is a loop starting at w which only uses

β-optimal actions. Define L_s^- to be the set of states $w \in L_s$ such that $v_\beta(w) \leq v_\beta(w')$ for every $w' \in L_s$. The reader can verify that the sets W_s, L_s, and L_s^- are all nonempty.

Let Θ_s be the collection of all pairs (h_s, ℓ_s) such that

- $h_s = (s_1, a_1, \ldots, s_m, a_m, s_{m+1})$ is a history that satisfies
  - $s_1 = s$ and $s_{m+1} \in L_s^-$, and the states $s_1, \ldots, s_{m+1}$ are all different,
  - action a_i is β-optimal in state s_i for all $i \in \{1, \ldots, m\}$.
- $\ell_s = (z_1, b_1, \ldots, z_k, b_k, z_{k+1})$ is a loop that satisfies
  - $z_1 = z_{k+1} = s_{m+1}$, and the states $z_1, \ldots, z_k$ are all different,
  - action b_i is β-optimal in state z_i for all $i \in \{1, \ldots, k\}$,
  - if ℓ_s visits a state that is also visited by h_s, then from this state ℓ_s follows h_s until reaching s_{m+1}. Formally, if $z_i = s_j$ for some $i \in \{2, \ldots, k\}$ and $j \in \{1, \ldots, m\}$, then

$$(b_i, z_{i+1}, b_{i+1}, \ldots, z_{k+1}) = (a_j, s_{j+1}, a_{j+1}, \ldots, s_{m+1}).$$

The set Θ_s is nonempty. Indeed, any path from s to a state in L_s^- that uses only β-optimal actions and any assignment of β-optimal actions to the states not on the path induce a pair in Θ_s. Moreover, each pair $(h_s, \ell_s) \in \Theta_s$ induces an infinite history $g(h_s, \ell_s)$, which starts at the state s as follows: Follow the history h_s until its final state, which is by definition the first state of the loop, and from then on follow the loop ℓ_s repeatedly. Note that this infinite history is stationary in the sense that, if a state is visited twice, the same action is being prescribed. The main point is that h_s ends in a state with a minimal discounted value in L_s. Due to optimality, this implies that the discounted sum of the payoffs along h_s is the highest among all paths that lead to L_s.

Take an enumeration $y_1, \ldots, y_n$ of the set of states S. We recursively define a sequence $(h_1, \ell_1) \in \Theta_{y_1}, \ldots, (h_n, \ell_n) \in \Theta_{y_n}$ as follows. Take an arbitrary $(h_1, \ell_1) \in \Theta_{y_1}$. Suppose that for some $k < n$, the sequence $(h_1, \ell_1) \in \Theta_{y_1}, \ldots, (h_k, \ell_k) \in \Theta_{y_k}$ has been defined. We define (h_{k+1}, ℓ_{k+1}). Take an arbitrary element $(h'_{k+1}, \ell'_{k+1}) \in \Theta_{y_{k+1}}$. We distinguish two cases:

Case 1: The pair (h'_{k+1}, ℓ'_{k+1}) only uses states that are not used by any of the pairs $(h_1, \ell_1), \ldots, (h_k, \ell_k)$. In this case, define $(h_{k+1}, \ell_{k+1}) = (h'_{k+1}, \ell'_{k+1})$.

Case 2: Assume that the pair (h'_{k+1}, ℓ'_{k+1}) uses a state that is used by at least one of the pairs $(h_1, \ell_1), \ldots, (h_k, \ell_k)$. In this case, let m be the smallest index such that (h'_{k+1}, ℓ'_{k+1}) uses a state that is used by (h_m, ℓ_m). Suppose that the first state of (h'_{k+1}, ℓ'_{k+1}) with this property is state z (here, the states of h'_{k+1} are considered before the states of ℓ'_{k+1}). Let w_m denote the final state of h_m, which is then also the first state of ℓ_m. Let h be the history that follows (h'_{k+1}, ℓ'_{k+1}) until reaching state z and then switches to (h_m, ℓ_m) until reaching state w_m. We argue that (h, ℓ_m) belongs to $\Theta_{y_{k+1}}$. All properties are easy to check. We only verify that $w_m \in L^-_{y_{k+1}}$. Let w'_{k+1} denote the final state of h'_{k+1}. By the existence of the common state z, we have $w'_{k+1} \in W_{y_m}$. Since $w'_{k+1} \in L_{y_{k+1}}$, we also obtain $w'_{k+1} \in L_{y_m}$. Therefore, by $(h_m, \ell_m) \in \Theta_{y_m}$, we derive $v_\beta(w_m) \leq v_\beta(w'_{k+1})$. Consequently, $w_m \in L^-_{y_{k+1}}$ as claimed. Let $(h_{k+1}, \ell_{k+1}) = (h, \ell_m)$.

Definition of σ^*: Now we define σ^* as follows: Starting with the initial state y_k, follow the infinite history $g(h_k, \ell_k)$. By construction, σ^* is well defined and stationary. The strategy σ^* is β-optimal, as it only uses β-optimal actions.

Proving that σ^* *is SO* β*-optimal.* Fix an arbitrary state $s = y_k$. Let $h^* = h_k$ and $\ell^* = h_k$ denote the corresponding history and loop, and let s^* denote the final state of h^*.

For simplicity, we assume w.l.o.g. that $v_\beta(s) = 0$. Let k denote the length of h^*, and let m denote the length of ℓ^*. Consider the infinite sequence of stages $E = \{k, k+m, k+2m, \ldots\}$. At these stages, the play according to σ^* is in state s^*. Let r_t^* denote the payoff at stage t with respect to σ^*. Because σ^* is β-optimal, we have for every $t \in E$

$$0 = v_\beta(s) = (1-\beta)\left(r_1^* + \beta r_2^* + \cdots + \beta^{t-2} r_{t-1}^*\right) + \beta^{t-1} v_\beta(s^*).$$

Hence, for every $t \in E$

$$v_\beta(s^*) = -\frac{1-\beta}{\beta^{t-1}}\left(r_1^* + \beta r_2^* + \cdots + \beta^{t-2} r_{t-1}^*\right). \tag{6}$$

Take any strategy σ, and let r_t and s_t denote the payoff and the state at stage t with respect to σ. We prove that there exists an infinite subset E' of E such that for every $t \in E'$

$$u_{\beta\, t-1}(s, \sigma^*) \geq u_{\beta\, t-1}(s, \sigma). \tag{7}$$

We can assume that σ is β-optimal; otherwise, this inequality holds for every sufficiently large stage $t \in E$. Since there are finitely many states, there are an infinite subset E' of E and a state $z \in S$ such that $s_t = z$ for every $t \in E'$. Because σ is β-optimal, it follows that $z \in L_s$.

For every stage $t \in E'$, similarly to (6), we have

$$v_\beta(z) = -\frac{1-\beta}{\beta^{t-1}}\left(r_1 + \beta r_2 + \cdots + \beta^{t-2} r_{t-1}\right).$$

Since $s^* \in L_s^-$, we have $v_\beta(s^*) \leq v_\beta(z)$, and therefore, for every $t \in E'$

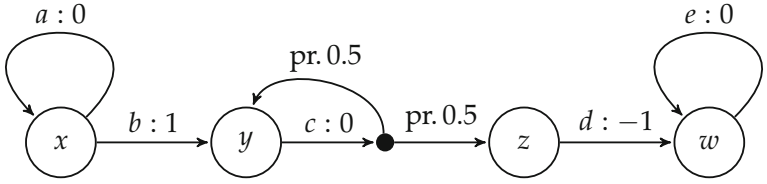
$$-\frac{1-\beta}{\beta^{t-1}}\left(r_1^* + \beta r_2^* + \cdots + \beta^{t-2} r_{t-1}^*\right) \leq -\frac{1-\beta}{\beta^{t-1}}\left(r_1 + \beta r_2 + \cdots + \beta^{t-2} r_{t-1}\right).$$

This implies (7) for every $t \in E'$, which proves that σ^* is SO β-optimal for the initial state y_k.

6 Nondeterministic MDPs

In this section, we describe two examples demonstrating that SO optimal strategies may fail to exist in nondeterministic MDPs.

Example 5 Consider the following MDP with four states, in which only one of the transitions is nondeterministic. The notation is as before, except for action c in state y that yields payoff 0 and from which transition is not deterministic: The process remains in state y with probability 0.5 and moves to state z with probability 0.5.



For every $n \in \mathbb{N} \cup \{\infty\}$, let σ_n be the strategy that chooses action a at all stages $t < n$ and chooses action b at stage n. For every horizon $T \geq n+2$, we have $u_T(x, \sigma_\infty) = 0$ and

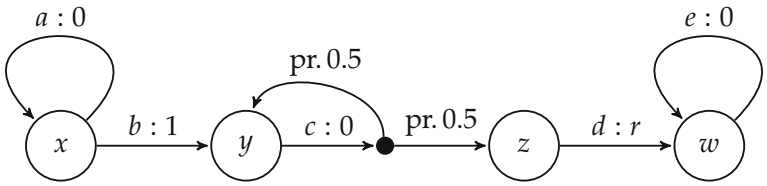
$$u_T(x, \sigma_n) = \frac{1}{T} 0.5^{T-n-1}.$$

It follows in particular that σ_∞ is not SO average optimal. Moreover, $u_T(x, \sigma_n) < u_T(x, \sigma_{n+1})$ for all $T \geq n+3$, implying that the strategy σ_n is not SO average optimal. Since this holds for each $n \in \mathbb{N}$, the MDP has no SO average optimal strategy.

The intuition for this result is as follows: The strategy σ_1 (playing action b at stage 1) induces a positive average expected payoff at each finite horizon T, due to the fact that the probability to remain in the state y up to stage T is positive. This payoff actually decays as the horizon increases, because state w is visited from state y in two moves with probability 0.5. Waiting for 1 period in the state x increases the probability of being in state y at any stage $T \geq 2$ and thus induces a higher average expected payoff at each finite horizon $T \geq 2$, than moving to the state y immediately.

We slightly modify the example above to produce an example where there is no SO discounted optimal strategy.

Example 6 Consider the following MDP with four states in which only one of the transitions is nondeterministic, and let $\beta \in [0, 1)$ be an arbitrary discount factor.



Here the payoff $r < 0$ is chosen so that the strategy σ_1 induces a β-discounted expected payoff equal to 0. Thus r satisfies the equation

$$u_\beta(x, \sigma_1) = (1-\beta)\left(1 + r(\beta^2 0.5 + \beta^3 0.5^2 + \cdots)\right) = 0.$$

Now for each $T \geq 3$ we have

$$\begin{aligned}(1 + \cdots + \beta^{T-1}) \cdot u_{\beta,T}(x, \sigma_1) &= 1 + r\beta \sum_{k=3}^{T} (0.5\beta)^{k-2} \\ &= -r\beta \sum_{k=T+1}^{\infty} (0.5\beta)^{k-2} \\ &= -\left(\frac{r\beta}{1 - 0.5\beta}\right)(0.5\beta)^{T-1}.\end{aligned}$$

The strategy σ_n could be thought of as implementing σ_1 after $n-1$ stages of delay. Hence for each $T \geq n+2$, we have

$$u_{\beta,T}(x, \sigma_n) = \beta^{n-1} u_{\beta,T-n+1}(x, \sigma_1).$$

Consequently

$$(1+\cdots+\beta^{T-1})\cdot u_{\beta,T}(x,\sigma_n) = -\left(\frac{r\beta}{1-0.5\beta}\right)\beta^{n-1}(0.5\beta)^{T-n}$$
$$= -\left(\frac{r\beta}{1-0.5\beta}\right)\beta^{T-1}0.5^{T-n}.$$

It follows in particular that $u_{\beta,T}(x,\sigma_n) < u_{\beta,T}(x,\sigma_{n+1})$ for each $T \geq n+3$, showing that σ_n is not SO β-optimal. Since this holds for each $n \in \mathbb{N}$, and since σ_∞ is not SO β-optimal, we conclude that the MDP has no SO β-optimal strategy.

7 Extensions

7.1 SO Optimality Based on Realized Payoffs

In the preceding section, we have seen that an SO optimal strategy may fail to exist in nondeterministic MDPs. In this section, we discuss two alternative definitions based on the comparison of realized, rather than expected, payoffs. We restrict ourselves to the case of average payoffs. Similar definitions could be given for the case of discounted payoffs.

Consider an MDP, possibly a nondeterministic one. We define a relation $\prec$ on strategies as follows: $\sigma' \prec \sigma$ if there exists $T \in \mathbb{N}$ such that for every $t \geq T$ the distribution of $r(s_1,a_1)+\cdots+r(s_t,a_t)$ induced by σ first order stochastically dominates the analog distribution induced by σ'. Notice that if $\sigma \prec \sigma'$, then $u_t(s,\sigma) < u_t(s,\sigma')$ for every $t \geq T$. The converse is true in deterministic MDPs but not in general. Consequently, the maximal elements of $\prec$ are exactly the SO optimal strategies when the underlying MDP is deterministic. In nondeterministic MDPs, every SO optimal strategy is maximal with respect to $\prec$, but not vice versa. Example 5 shows that a maximal element of $\prec$ need not exist. Indeed, in this example the sum of the payoffs can only be 0 or 1, so that one distribution of total payoffs dominates another when it has a higher expected utility.

We turn to yet another definition. We write $\sigma' \prec_\infty \sigma$ if there are sets $P \subseteq H_\infty$ and $P' \subseteq H_\infty$ satisfying the following two properties: [1] With probability 1, σ induces an infinite history in P, and with probability 1, σ' induces an infinite history in P'; [2] for any infinite histories $(s_1,a_1,s_2,a_2,\ldots) \in P$ and $(s'_1,a'_1,s'_2,a'_2,\ldots) \in P'$, there is a horizon T such that for all $t \geq T$

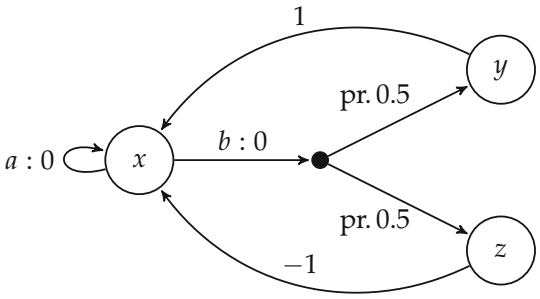
$$r(s_1,a_1)+\cdots+r(s_t,a_t) > r(s'_1,a'_1)+\cdots+r(s'_t,a'_t).$$

The maximal elements of $\prec_\infty$ are the SO optimal strategies when the underlying MDP is deterministic. In nondeterministic MDPs, maximality with respect to $\prec_\infty$ is unrelated to SO optimality, as we argue below.

On the one hand, it is easy to see that maximality with respect to $\prec_\infty$ does not imply SO optimality. Indeed, consider an MDP with three absorbing states x_2, x_1, and x_{-1}, with the payoffs 2, 1, and -1 respectively. There is also state x_0 with two possible actions. Action a leads with probability 0.5 to x_2, and with probability 0.5 to x_{-1}. Action b leads with probability 0.5 to x_1, and with probability 0.5 to x_{-1}. Both actions are maximal with respect to $\prec_\infty$, but only a is SO optimal.

On the other hand, SO optimality does not imply maximality with respect to $\prec_\infty$. Intuitively, under $\prec_\infty$ it is very difficult to beat a strategy. It is therefore surprising that a maximal element with respect to $\prec_\infty$ may fail to exist, as the example below demonstrates.

Example 7 Consider the following MDP with three states and one nondeterministic transition. In state x of this MDP, action a yields payoff 0 and the process remains in state x, whereas action b first yields payoff 0 and subsequently payoff 1 with probability 0.5 and payoff -1 with probability 0.5, before returning to state x.



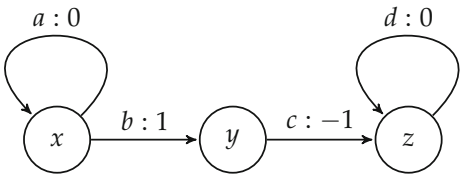
In this MDP, all strategies are SO average optimal. The relation $\prec_\infty$ has no maximal element. Indeed, take any $n \in \mathbb{N}$ and consider the following strategy σ_n: Play action b until the sum of the payoffs reaches n, and play action a in all subsequent stages. The crucial feature of this strategy is that the sum of the payoffs eventually reaches n with probability 1. The reason is that the sum of the payoffs is a symmetric random walk on the integers, as long as action b is played, and it is well known that a symmetric random walk eventually reaches any given integer value with probability 1. It is clear that $\sigma_n \prec_\infty \sigma_{n+1}$. Since this is true for each n, the relation $\prec_\infty$ has no maximal element.

7.2 SO Optimality in Randomized Strategies

We briefly examine SO optimality when the decision maker may use randomized strategies. A randomized strategy τ is a map that, to each history $h \in H$, assigns a probability distribution $\tau(h)$ on the set of available actions $A(s(h))$. The interpretation is that τ recommends to choose an action according to $\tau(h)$ if history h arises.

We show that deterministic MDPs do not always admit SO optimal strategies when randomized strategies are allowed, so Theorem 2 does not extend to randomized strategies.

Example 8 Consider the following MDP with three states, in which the initial state is x.



Every pure strategy is SO average optimal. Indeed, for every pure strategy σ, the sum of the payoffs is 0 for sufficiently large horizons: $u_T(x, \sigma) = 0$ for large $T \in \mathbb{N}$. Now consider the randomized strategy τ_p, for $p \in (0, 1)$, that in state x always chooses action a with probability p and action b with probability $1 - p$. With respect to τ_p, for any $p \in (0, 1)$, the expected sum of the payoffs up to any stage T is positive: $u_T(x, \tau_p) > 0$ for all T. So, in the SO sense, any τ_p is better than any pure strategy. In fact, $\tau_{p'}$ is also better than τ_p if

$0 < p < p' < 1$. In conclusion, this MDP does not admit SO average optimal strategies when randomized strategies are allowed.

An example that shows that Theorem 2 does not extend to randomized strategies for SO discounted optimality can be constructed along similar lines.

8 Concluding Remarks

In Sect. 3, we provided two different proofs for the second part of Theorem 2 on the existence of a stationary strategy that is SO average optimal for all initial states. Both proofs make use of results for the discounted evaluation.

The question naturally arises whether there is a way to construct SO average optimal strategies without the use of the discounted game. The SO discounted optimal strategy constructed in the proof of Theorem 2 first follows a path h and then repeats a loop ℓ ad infinitum. The path h and loop ℓ are chosen such that (a) the pair (h, ℓ) maximizes the discounted payoff, and (b) subject to part (a), it maximizes the discounted payoff along the path h.

In view of Theorem 1, the above strategy is also SO average optimal if the discount factor is high enough. Thus, the path h and loop ℓ have the property that (a') the pair (h, ℓ) (or equivalently the loop ℓ) maximizes the average payoff, and (b') subject to part (a'), it maximizes the sum of payoffs along the path h.

The converse is not necessarily true: It can be the case that a pair (h', ℓ') of a path and a loop satisfies (a') and (b'), but they do not induce a strategy that is SO discounted optimal for large discount factor. Indeed, as an example, suppose that ℓ' induces the sequence of payoffs $(0, 0, \dots)$, while there is another loop with the same starting state that induces the sequence of payoffs $(1, -1, 1, -1, \dots)$. Then ℓ' is not discounted optimal for any discount factor.

However, one can prove that if the pair (h', ℓ') satisfies properties (a') and (b'), then the induced strategy is SO optimal for the corresponding initial state. Notice that this construction does not make use of the discounted payoffs.

References

1. Blackwell D (1962) Discrete dynamic programming. Ann Math Stat 33:719–726
2. Carlson DA, Haurie A, Leizarowitz A (1991) Infinite horizon optimal control. Springer, Berlin
3. Guo X, Hernández-Lerma O (2009) Continuous-time Markov decision processes. Springer, Berlin
4. Howard RA (1960) Dynamic programming and markov processes. MIT Press, Cambridge
5. Leizarowitz A (1996) Overtaking and almost-sure optimality for infinite Horizon Markov decision processes. Math Oper Res 21:158–181
6. Méder Z, Flesch J, Peeters R (2012) Optimal choice for finite and infinite horizons. Oper Res Lett 40:469–474
7. Nowak AS, Vega-Amaya O (1999) A counterexample on overtaking optimality. Math Methods Oper Res 49:435–439
8. Puterman ML (1994) Markov decision processes: discrete stochastic dynamic programming. Wiley, New York
9. Shapley LS (1953) Stochastic games. Proc Natl Acad Sci 39:1095–1100
10. Stern LE (1984) Criteria of optimality in the infinite-time optimal control problem. J Optim Theory Appl 44:497–508

11. Zaslavski AJ (2006) Turnpike properties in the calculus of variations and optimal control. Springer, New York
12. Zaslavski AJ (2014) Turnpike phenomenon and infinite horizon optimal control. Springer Optimization and Its Applications, New York