# *A Counterexample on Sample-Path Optimality in Stable Markov Decision Chains with the Average Reward Criterion*

**Rolando Cavazos-Cadena, Raúl Montes-de-Oca & Karel Sladký**

**Journal of Optimization Theory and Applications**

ISSN 0022-3239
Volume 163
Number 2

J Optim Theory Appl (2014) 163:674-684
DOI 10.1007/s10957-013-0474-6


Springer

J Optim Theory Appl (2014) 163:674–684
DOI 10.1007/s10957-013-0474-6

# A Counterexample on Sample-Path Optimality in Stable Markov Decision Chains with the Average Reward Criterion

**Rolando Cavazos-Cadena · Raúl Montes-de-Oca · Karel Sladký**

Received: 22 September 2012 / Accepted: 30 October 2013 / Published online: 23 November 2013


**Abstract** This note deals with Markov decision chains evolving on a denumerable state space. Under standard continuity-compactness requirements, an explicit example is provided to show that, with respect to a strong sample-path average reward criterion, the Lyapunov function condition does not ensure the existence of an optimal stationary policy.

**Keywords** Strong sample-path optimality · Lyapunov function condition · Stationary policy · Expected average reward criterion

## 1 Introduction

This work is concerned with discrete-time Markov decision processes (MDPs) with denumerable state space, compact action sets, and endowed with a long-run average criterion. It is assumed that the reward function and the transition law depend continuously on the applied action, and that the so-called Lyapunov function condition holds. Within this framework, the optimal expected average reward does not depend

R. Cavazos-Cadena (✉)
Departamento de Estadística y Cálculo, Universidad Autónoma Agraria Antonio Narro, Buenavista, Saltillo Coah 25315, Mexico
e-mail: rcavazos@uaaan.mx

R. Montes-de-Oca
Departamento de Matemáticas, Universidad Autónoma Metropolitana, Campus Iztapalapa, Avenida San Rafael Atlixco 186, Colonia Vicentina, México 09340, Mexico
e-mail: momr@xanum.uam.mx

K. Sladký
Institute of Information Theory and Automation, Pod Vodárenskou věží 4, 182 08 Praha 8, Czech Republic
e-mail: sladky@utia.cas.cz

on the initial state and is characterized in terms of the corresponding optimality equation, whose solution renders an *expected* average optimal stationary policy; see [1]. The expected average reward index arises naturally when the controller runs the underlying dynamical system a large number of times, but such an expected criterion does not look appropriate if the system is going to be observed just a few times. In this latter case, it is natural to analyze the behavior of a control policy from a *sample-path perspective*, and the following notion will be used in this note:

> A policy $\pi^*$ is average optimal in the sample-path sense if there exists a constant, say $g^*$, such that under the action of $\pi^*$ and regardless the initial state, the average of the observed rewards over a finite horizon $t$ converges to $g^*$ as $t \to \infty$ with probability 1, whereas under any other policy the superior limit of such averages is always bounded above by $g^*$ almost surely.

Under the mild continuity-compactness conditions in this note, it was recently shown in [2] that if the existence of a Lyapunov function is complemented with an additional requirement, which is described in Sect. 3, then an expected average optimal stationary policy obtained from the optimality equation is also sample-path average optimal in the sense described above, a result that naturally leads to consider the following question:

- Can the existence of a sample-path optimal stationary policy be ensured under the sole assumption that the decision model admits a Lyapunov function?

*The main objective* of this work is to exhibit an explicit example showing that the answer to this question is negative, a conclusion that establishes an interesting contrast between the expected and sample-path average criteria. *The organization* of the subsequent material is as follows. In Sect. 2 the decision model and the Lyapunov function condition are briefly discussed, whereas in Sect. 3 the notion of (strong) sample-path average optimal policy and the main question considered in this note are formally stated. Next, in Sect. 4 an MDP admitting a Lyapunov function and satisfying the standard continuity-compactness requirements is introduced and, for such a particular model, it is shown that there is not any stationary policy which is sample-path average optimal in the strong sense of Definition 3.1. Finally, the exposition concludes in Sect. 5 with some brief comments.

## 2 Decision Model

Throughout the remainder $\mathbb{N}$ stands for the set of all nonnegative integers and the indicator function of an event $A$ is denoted by $I[A]$; given sequence of events $\{A_n\}$, the corresponding superior limit is denoted by $[A_n \text{ i.o.}]$, that is, $[A_n \text{ i.o.}] := \bigcap_{m=1}^{\infty} \bigcup_{k=m}^{\infty} A_k$, whereas the class of all real-valued and continuous functions defined on a topological space $\mathbb{K}$ is denoted by $\mathcal{C}(\mathbb{K})$. Now, let $\mathcal{M} = (S, A, \{A(x)\}_{x \in S}, R, P)$ be an MDP, where the state space $S$ is a denumerable set endowed with the discrete topology, the action set $A$ is a metric space and, for each $x \in S$, $A(x) \subset A$ is the nonempty subset of admissible actions at $x$, whereas $R \in \mathcal{C}(\mathbb{K})$ is the reward function, where $\mathbb{K} := \{(x, a) | x \in S, a \in A(x)\}$ is the space

of admissible pairs. On the other hand, $P = [p_{xy}(\cdot)]$ is the controlled transition law on $S$ given $\mathbb{K}$, that is, for all $(x, a) \in \mathbb{K}$ and $y \in S$, the relations $p_{xy}(a) \geq 0$ and $\sum_{y \in S} p_{xy}(a) = 1$ are satisfied. This model $\mathcal{M}$ is interpreted as follows: At each time $t \in \mathbb{N}$ the decision maker knows the previous states and actions and observes the current state, say $X_t = x \in S$. Using that information, the controller selects an action (control) $A_t = a \in A(x)$ and two things happen: a reward $R(x, a)$ is obtained by the controller, and the system moves to a new state $X_{t+1} = y \in S$ with probability $p_{xy}(a)$.

**Assumption 2.1**

(i) For each $x \in S$, $A(x)$ is a compact subset of $A$.
(ii) For every $x, y \in S$, the mappings $a \mapsto R(x, a)$ and $a \mapsto p_{xy}(a)$ are continuous in $a \in A(x)$.

*Policies* A policy $\pi$ is a (measurable) rule for choosing actions which, at each time $t \in \mathbb{N}$, may depend on the current state and on the record of previous states and actions; see, for instance, [3] for details. The class of all policies is denoted by $\mathcal{P}$ and, given the initial state $x \in S$ and the policy $\pi$ being used for choosing actions, the distribution of the state-action process $\{(X_t, A_t)\}$ is uniquely determined; such a distribution is denoted by $P_x^\pi$, whereas $E_x^\pi$ stands for the corresponding expectation operator. Next, define $\mathbb{F} := \prod_{x \in S} A(x)$ and notice that $\mathbb{F}$ is a compact metric space, which consists of all functions $f : S \to A$ such that $f(x) \in A(x)$ for each $x \in S$. A policy $\pi$ is *stationary* iff there exists $f \in \mathbb{F}$ such that the equality $A_t = f(X_t)$ is always valid under $\pi$; in this case $\pi$ and $f$ are naturally identified and, with this convention, $\mathbb{F} \subset \mathcal{P}$.

*Expected Average Criterion and Lyapunov Function Condition* Assume that $R(X_t, A_t)$ has finite expectation with respect to every distribution $P_x^\pi$. The (long-run superior limit) expected average reward criterion corresponding to $\pi \in \mathcal{P}$ at state $x \in S$ is defined by

$$J(x, \pi) := \limsup_{k \to \infty} \frac{1}{k} E_x^\pi \left[ \sum_{t=0}^{k-1} R(X_t, A_t) \right], \tag{1}$$

whereas the corresponding optimal value function is

$$J^*(x) := \sup_{\pi \in \mathcal{P}} J(x, \pi), \quad x \in S; \tag{2}$$

a policy $\pi^* \in \mathcal{P}$ is (expected) average optimal if $J(x, \pi^*) = J^*(x)$ for every $x \in S$. A fundamental instrument to analyze the above criterion is the following *optimality equation*:

$$g + h(x) = \sup_{a \in A(x)} \left[ R(x, a) + \sum_{y \in S} p_{xy}(a) h(y) \right], \quad x \in S, \tag{3}$$

where $g \in \mathbb{R}$ and $h \in \mathcal{C}(S)$ is a given function. Suppose that the pair $(g, h(\cdot))$ satisfies (3) and that the following properties are valid: For each $x \in S$ and $\pi \in \mathcal{P}$,

$$\text{(i)} \quad E_x^\pi\big[|h(X_n)|\big] < \infty \text{ for each } n \in \mathbb{N}, \text{ and } E_x^\pi\big[|h(X_n)|\big]/n \to 0 \text{ as } n \to \infty;$$

$$\text{(ii)} \quad \text{The mapping } a \mapsto \sum_{y \in S} p_{xy}(a)h(y), \ a \in A(x) \text{ is continuous.} \tag{4}$$

Using Assumption 2.1, these requirements yield that

(a) $J^*(x) = g$ for each $x \in S$, and
(b) There exists a stationary policy $f \in \mathbb{F}$ satisfying

$$g + h(x) = R\big(x, f(x)\big) + \sum_{y \in S} p_{xy}\big(f(x)\big)h(y), \quad x \in S, \tag{5}$$

and such a stationary policy $f$ is average optimal [1, 2].

The existence of a solution $(g, h(\cdot))$ of the optimality equation satisfying the properties (i) and (ii) in (4) and, consequently, rendering the above conclusions (a) and (b), requires some connectedness condition [4]. The following is a general requirement in this direction.

**Assumption 2.2** (Lyapunov Function Condition [1]) There exists $z \in S$ and a function $\ell\colon S \to [1, \infty)$ satisfying the properties (i)–(iii) below:

(i) $1 + |R(x, a)| + \sum_{y \neq z} p_{xy}(a)\ell(y) \leq \ell(x)$ for all $(x, a) \in \mathbb{K}$;
(ii) For each $x \in S$, $a \mapsto \sum_y p_{xy}(a)\ell(y)$ is a continuous function of $a \in A(x)$;
(iii) For every $f \in \mathbb{F}$ and $x \in S$, $E_x^f[\ell(X_n)I[T > n]] \to 0$ as $n \to \infty$, where

$$T := \min\big\{n > 0 | X_n = z\big\}$$

is the first return time to state $z$.

Given a state $z \in S$, a function $\ell$ satisfying the conditions (i)–(iii) in the above assumption is referred to as a *Lyapunov function* for the model $\mathcal{M}$.

**Lemma 2.1** [1] *Under Assumptions 2.1 and 2.2, assertions* (i)–(iii) *below hold*:

(i) $E_x^\pi[|R(X_n, A_n)|]$ *is finite for every* $x \in S$ *and* $\pi \in \mathcal{P}$;
(ii) *There exists a pair* $(g, h(\cdot)) \in \mathbb{R} \times \mathcal{C}(S)$ *satisfying the optimality equation* (3) *as well as the two conditions in* (4);
(iii) *There exists a policy* $f \in \mathbb{F}$ *satisfying* (5) *and such a stationary policy is expected average optimal*; *moreover*, $J^*(\cdot) = g$.

## 3 Sample-Path Optimality

The expected average criterion in (1) is quite appropriate if the controller repeats the underlying random dynamical experiment many times under similar conditions, but

not for a single trial. In this latter case, it is interesting to study the average reward from a *sample-path* point of view.

**Definition 3.1** A policy $\pi^* \in \mathcal{P}$ is *(strong)* sample-path average optimal with optimal value $g^* \in \mathbb{R}$ if the following conditions (i) and (ii) hold:

(i) For each state $x \in S$, $\lim_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} R(X_t, A_t) = g^*$ $P_x^{\pi^*}$-a.s., and

(ii) For every $\pi \in \mathcal{P}$ and $x \in S$, $\limsup_{n\to\infty} \frac{1}{n} \sum_{t=0}^{n-1} R(X_t, A_t) \leq g^*$ $P_x^{\pi}$-a.s..

The above notion is stronger than the idea of sample path optimality that was employed in [5], where the second property in the above definition was replaced by the weaker requirement $\liminf_{n\to\infty} \sum_{t=0}^{n-1} R(X_t, A_t)/n \leq g^*$ $P_x^{\pi}$-a.s.. *With respect to this weaker notion of sample-path optimality*, it was proved in Theorem 4.1 of the aforementioned paper that the policy $f \in \mathbb{F}$ satisfying (5) is sample-path optimal with optimal sample-path average reward $g^* = J^*(\cdot)$. On the other hand, it was recently shown in [2] that, under Assumptions 2.1 and 2.2, if the Lyapunov function $\ell$ is such that, regardless of the initial state and the policy employed, the expected average reward corresponding to $\ell^\beta$ is finite for some $\beta > 2$, then the stationary policy $f$ in (5) is sample-path average optimal in the strong sense. In short, the following results are presently available for MDPs satisfying Assumptions 2.1 and 2.2: (a) There exists a stationary policy which is sample-path optimal with respect to a criterion that is *weaker* than the one specified above; and (b) With respect to the strong idea of sample-path optimality in Definition 3.1, the existence of an optimal stationary policy has been established when the Lyapunov function satisfies *an additional requirement.* These facts naturally lead to the following question:

> Are Assumptions 2.1 and 2.2 sufficient to ensure the existence of a stationary policy that is sample-path optimal in the sense of Definition 3.1?

*The main contribution of this note* consists in showing that the answer to this question is negative, a conclusion that will be established using an explicit example; for additional results on sample-path optimality in other contexts see, for instance, [6] and [7].

## 4 The Counterexample

In this section an example will be given to show that Assumptions 2.1 and 2.2 do not generally ensure the existence of a sample-path optimal stationary policy in the sense of Definition 3.1.

*Example 4.1* Let the state space $S$ and the action set $A$ be the topological subspaces of the real line given by

$$S = \mathbb{N}, \quad A = \{0\} \cup \big\{1/k | k = 1, 2, 3, \ldots\big\},$$

and define the sets of admissible actions by

$$A(0) = A, \quad \text{and} \quad A(x) = \{1\}, \quad x = 1, 2, 3, \ldots.$$

Next, let the transition law be determined by

$$p_{x0}(1) = 1, \quad x = 1, 2, 3, \ldots, \tag{6}$$

$$p_{00}(0) = 1, \quad \text{and} \quad p_{0x}(1/x) = \frac{1}{x \log(12 + x)} = 1 - p_{00}(1/x), \quad x = 1, 2, \ldots \tag{7}$$

and, finally, let the reward function $R\colon \mathbb{K} \to \mathbb{R}$ be given by

$$R(x, a) = ax, \quad x \in S, \quad a \in A(x), \tag{8}$$

that is, $R(0, a) = 0$ for every $a \in A(0) = A$, and $R(x, 1) = x$ for every $x \in S$.

It is not difficult to see that the continuity-compactness conditions in Assumption 2.1 are satisfied for the model in the above example, and it will be shown below that Assumption 2.2 also holds. First, using that $A(x)$ is the singleton $\{1\}$ for each state $x \neq 0$, observe that

$$\mathbb{F} = \{f_0, f_1, f_2, \ldots\}, \tag{9}$$

where the policies $f_k$ are determined by $f_k(x) = 1$ for every $x = 1, 2, 3, \ldots$ and $k \in \mathbb{N}$, and

$$f_0(0) = 0, \quad f_k(0) = 1/k, \quad k = 1, 2, 3, \ldots. \tag{10}$$

**Proposition 4.1** *Consider the model $\mathcal{M}$ in Example 4.1, set $z = 0$ and define the function $\ell\colon S \to \mathbb{R}$ as follows*:

$$\ell(0) = 1 + \frac{2}{\log(12)}, \quad \textit{and} \quad \ell(x) = x + 1, \quad x = 1, 2, 3, \ldots. \tag{11}$$

*With this notation, the following assertions* (*a*) *and* (*b*) *hold*:

(a) *The mapping $\ell$ is a Lyapunov function for the model $\mathcal{M}$*;
(b) *The optimal* expected *average reward is*

$$g = \frac{1}{\log(16) + 1/4}, \tag{12}$$

*and $f_4$ is the unique* (*expected*) *average optimal stationary policy*; *see* (10).

*Proof* (a) The three requirements in Assumption 2.2 will be verified.

(i) Note that (6)–(8) and (11) together yield that

$$1 + \big|R(x, 1)\big| + \sum_{y \neq 0} p_{xy}(1)\ell(y) = 1 + x = \ell(x), \quad x = 1, 2, 3, \ldots,$$

$$1 + \big|R(0, 0)\big| + \sum_{y \neq 0} p_{0y}(0)\ell(y) = 1 \leq \ell(0), \quad \text{and}$$

$$1 + \big|R(0, 1/k)\big| + \sum_{y \neq 0} p_{0y}(1/k)\ell(y) = 1 + \frac{k+1}{k\log(12+k)}$$
$$\leq 1 + \frac{2}{\log(12)} = \ell(0), \quad k = 1, 2, 3, \ldots;$$

from the specification of the action sets, these relations yield that the first condition in Assumption 2.2 is satisfied by $\ell$.

(ii) Since $A(x)$ is a singleton for $x \neq 0$ and zero is the unique accumulation point of $A(0) = \{0\} \cup \{1/k | k = 1, 2, 3, \ldots\}$, it is sufficient to show that

$$\lim_{k\to\infty} \sum_{y\in S} p_{0y}(1/k)\ell(y) = \sum_{y\in S} p_{0y}(0)\ell(y) = \ell(0),$$

a convergence that is valid, since (6), (7) and (11) together imply that

$$\sum_{y\in S} p_{0y}(1/k)\ell(y) = \frac{k+1}{k\log(12+k)} + \left(1 - \frac{1}{k\log(12+k)}\right)\ell(0), \quad k = 1, 2, 3, \ldots.$$

(iii) The specification of the transition law yields that, for every $x \in S$ and $f \in \mathbb{F}$, the inequality $T \leq 2$ holds $P_x^f$-a.s., so that $E_x^f[\ell(X_n)I[T > n]] = 0$ for $n \geq 2$.

(b) Using the existence of a Lyapunov function established in the previous part, Lemma 2.1 yields that the optimal expected average reward is constant, say $g$, and that

$$g = \sup_{f\in\mathbb{F}} J(0, f) = \sup\big\{J(0, f_k) | k = 0, 1, 2, 3, \ldots\big\}, \tag{13}$$

where (9) was used to set the second equality. Now, observe that under $f_0$ the state 0 is absorbing and then, since $R(0, 0) = 0$, it follows that

$$J(0, f_0) = 0. \tag{14}$$

Suppose now that the system is driven by policy $f_k$ with $k > 0$. In this case, the specification of the transition law yields that $X_1 = 0$ or $X_1 = k$ with probability 1, and the following assertions hold:

(i) On the event $[X_1 = 0]$ the equalities $T = 1$ and $\sum_{t=0}^{T-1} R(X_t, A_t) = 0$ are valid $P_0^{f_k}$-a.s.;
(ii) On $[X_1 = k]$, the events $[T = 2]$ and $[\sum_{k=0}^{T-1} R(X_k, A_k) = 0 + k = k]$ occur $P_0^{f_k}$-a.s.

These facts and the relations $P_0^{f_k}[X_1 = k] = 1/[k\log(12+k)] = 1 - P_0^{f_k}[X_1 = 0]$ together yield, via the theory of renewal-reward processes [8], that

$$J(0, f_k) = \frac{E_0^{f_k}[\sum_{k=0}^{T-1} R(X_k, A_k)]}{E_0^f[T]}$$
$$= \frac{k(1/[k\log(12+k)])}{(1 - 1/[k\log(12+k)]) + 2(1/[k\log(12+k)])}$$

$$= \frac{1/\log(12+k)}{1+1/[k\log(12+k)]}$$
$$= \frac{1}{\log(12+k)+1/k} \leq \frac{1}{\log(16)+1/4} = J(0, f_4),$$

where the inequality is strict for $k \neq 4$. The conclusion follows combining this last display with (13) and (14). □

If $(g, h(\cdot))$ is a solution of the optimality equation for the MDP in Example 4.1, then $f_4$ is the unique stationary policy satisfying (5), where the optimal expected average reward $g$ is given in (12). Also, notice that the standard ergodic theorem for Markov chains yields that, for every initial state $x \in S$ and $k = 0, 1, 2, 3, \ldots$ (see [8]),

$$\lim_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} R(X_t, A_t) = J(0; f_k) \leq g = \frac{1}{\log(16)+1/4} < \frac{1}{2}, \quad P_x^{f_k}\text{-a.s.} \quad (15)$$

This relation will be used to show that a sample-path average optimal stationary policy in the sense of Definition 3.1 does not exist. To achieve this goal, it is convenient to introduce some notation: For each $t \in \mathbb{N}$ let $N_t$ be the number of visits to state 0 up to time $t$, i.e.,

$$N_t(X_0, A_0, \ldots, X_{t-1}, A_{t-1}, X_t) \equiv N_t : = \sum_{k=0}^{t} I[X_k = 0]. \quad (16)$$

Using that $A(w) = \{1\}$ when $w \in S \setminus \{0\}$, the Markov property yields that, for every $x \in S$, $\delta \in \mathcal{P}$ and $t \in \mathbb{N}$,

$$P_x^{\delta}\big[X_t \neq 0, X_{t+1} \neq 0 | X_0, A_0, \ldots, X_t\big] = I[X_t \neq 0] \sum_{y \in S\setminus\{0\}} p_{X_t y}(1) = 0,$$

where (6) was used to set the second equality. Thus,

$$P_x^{\delta}[X_t \neq 0, X_{t+1} \neq 0] = 0, \quad (17)$$

i.e., $P_x^{\delta}[X_t = 0 \text{ or } X_{t+1} = 0] = 1$, an equality that via (16) yields the following conclusion.

**Proposition 4.2** *For each initial state $x \in S$ and $\delta \in \mathcal{P}$,*

$$P_x^{\delta}\big[N_t \geq [t/2]\big] = 1, \quad t = 0, 1, 2, 3, \ldots.$$

Next, let the policy $\pi \in \mathcal{P}$ be determined as follows: At each time $t \in \mathbb{N}$, under the action of $\pi$ the control $A_t$ is given by

$$A_t = 1/N_t \quad \text{if } X_t = 0, \quad \text{and} \quad A_t = 1 \quad \text{if } X_t \neq 0. \quad (18)$$

**Proposition 4.3** *The policy $\pi$ determined by (18) satisfies the following properties: For each $x \in S$,*

(i) $P_x^\pi[X_t \neq 0 \ i.o.] = 1$,

*and then*

(ii) $P_x^\pi[X_t = 0, X_{t+1} \neq 0 \ i.o.] = 1$.

*Proof* For each integer $m \in \mathbb{N}$, the specification of the policy $\pi$ yields that

$$\begin{aligned} &P_x^\pi\left[\bigcap_{k=m}^{\infty}[X_k = 0]|X_s, 0 \le s \le m\right] \\ &= I[X_m = 0]\prod_{k=0}^{\infty} p_{00}\big(1/[k + N_m]\big) \\ &= I[X_m = 0]\prod_{k=0}^{\infty}\left(1 - \frac{1}{(k + N_m)\log(12 + k + N_m)}\right) = 0, \end{aligned}$$

where the last equality is due to the relation $\sum_{r=1}^{\infty} 1/[r\log(12 + r)] = \infty$. Thus,

$$P_x^\pi\left[\bigcap_{k=m}^{\infty}[X_k = 0]\right] = 0, \quad x \in S, \quad m = 0, 1, 2, 3, \ldots,$$

a property that is equivalent to the first conclusion. Next, note that

$$P_x^\pi\big[[X_{t+1} \neq 0] \setminus [X_t = 0, X_{t+1} \neq 0]\big] = P_x^\pi[X_t \neq 0, X_{t+1} \neq 0] = 0,$$

where the second equality is due to (17); so, assertion (ii) follows from part (i). □

To conclude, the two previous propositions will be used to show that the MDP in Example 4.1 does not admit a sample path optimal stationary policy in the sense of Definition 3.1.

**Proposition 4.4** *In Example 4.1 the assertions* (i) *and* (ii) *below are valid*:

(i) *The policy* $\pi$ *in* (18) *satisfies the following property*: *For each* $x \in S$,

$$\limsup_{t\to\infty} \frac{\sum_{k=0}^{t-1} R(X_k, A_k)}{t} \ge \frac{1}{2} \quad P_x^\pi\text{-a.s.}$$

*Consequently*,

(ii) *A sample-path average optimal stationary policy does not exist*.

*Proof* (i) Let $x \in S$ be arbitrary and observe that (7) and (18) together imply that

$$\begin{aligned} P_x^\pi\big[X_t = 0, X_{t+1} \neq 0|X_0, X_1, \ldots X_t\big] &= I[X_t = 0]\sum_{y\neq 0} p_{0y}(1/N_t) \\ &= I[X_t = 0]p_{0N_t}(1/N_t) \\ &= P_x^\pi\big[X_t = 0, X_{t+1} = N_t|X_0, X_1, \ldots X_t\big], \end{aligned}$$

so that $P_x^\pi[X_t = 0, X_{t+1} \neq 0] = P_x^\pi[X_t = 0, X_{t+1} = N_t]$; combining this equality with the inclusion

$$[X_t = 0, X_{t+1} = N_t] \subset [X_t = 0, X_{t+1} \neq 0]$$

it follows that the events $[X_t = 0, X_{t+1} \neq 0]$ and $[X_t = 0, X_{t+1} = N_t]$ differ by a null set with respect to $P_x^\pi$, and then

$$P_x^\pi[X_t = 0, X_{t+1} = N_t, \text{i.o.}] = 1,$$

by Proposition 4.3(ii). Next, observe that the specification of the nonnegative reward function in (8), yields that

$$[X_t = 0, X_{t+1} = N_t] \subset \big[R(X_{t+1}, A_{t+1}) = N_t\big] \subset \left[\frac{\sum_{k=0}^{t+1} R(X_k, A_k)}{t+2} \geq \frac{N_t}{t+2}\right],$$

a relation that combined with the previous display leads to

$$P_x^\pi\left[\frac{\sum_{k=0}^{t+1} R(X_k, A_k)}{t+2} \geq \frac{N_t}{t+2} \text{ i.o.}\right] = 1,$$

and part (i) follows from this equality via Proposition 4.2.

(ii) Assume that $f_k \in \mathbb{F}$ is sample-path average optimal with optimal value $g^*$. In this case, the first part of Definition 3.1 and (15) together imply that $g^* \leq g$, and then the second condition in Definition 3.1 applied to the policy $\pi$ in (18) yields that, for every $x \in S$,

$$\limsup_{n\to\infty} \frac{1}{n+1} \sum_{k=0}^{n} R(X_t, A_t) \leq g^* \leq g \quad P_x^\pi\text{-a.s.},$$

a statement that, since $g < 1/2$, contradicts part (i). Therefore, a sample-path average optimal stationary policy in the sense of Definition 3.1 does not exist. □

## 5 Conclusion

This work considered discrete-time Markov decision chains on a denumerable state space. Besides standard continuity-compactness requirements, the main feature of the models analyzed in this note is that they admit a Lyapunov function. As already noted, in that context there exists a stationary policy which is optimal with respect to the expected average reward index; however, it was shown in Sect. 4 that Assumptions 2.1 and 2.2 do not imply the existence of a sample-path optimal stationary policy as specified in Definition 3.1, a conclusion that signals an interesting contrast between the expected and sample-path perspectives to the average criterion.

**Acknowledgements** This work was supported in part by the PSF Organization under Grant No. 012/300/02, and by CONACYT (México) and ASCR (Czech Republic) under Grant No. 171396.

The authors are grateful to the editor for helpful suggestions.

## References


1. Hordijk, A.: Dynamic Programming and Potential Theory. Mathematical Centre Tract, vol. 51. Mathematisch Centrum, Amsterdam (1974)
2. Cavazos-Cadena, R., Montes-de-Oca, R.: Sample-path optimality in average Markov decision chains under a double Lyapunov function condition. In: Hernández-Hernández, D., Minjárez-Sosa, A. (eds.) Optimization, Control, and Applications of Stochastic Systems, In Honor of Onésimo Hernández-Lerma, pp. 31–57. Springer, New York (2012)
3. Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley, New York (1994)
4. Thomas, L.C.: Connectedness conditions for denumerable state Markov decision processes. In: Hartley, R., Thomas, L.C., White, D.J. (eds.) Recent Developments in Markov Decision Processes, pp. 181–204. Academic Press, London (1980)
5. Cavazos-Cadena, R., Fernández-Gaucherand, E.: Denumerable controlled Markov chains with average reward criterion: sample path optimality. Math. Methods Oper. Res. **41**, 89–108 (1995)
6. Lasserre, J.B.: Sample-path average optimality for Markov control processes. IEEE Trans. Autom. Control **44**, 1966–1971 (1999)
7. Hunt, F.Y.: Sample path optimality for a Markov optimization problems. Stoch. Process. Appl. **115**, 769–779 (2005)
8. Ross, S.M.: Applied Probability Models with Optimization Applications. Holden-Day, Oakland (1970)