Another set of conditions for Markov decision processes with average sample-path costs

Quanxin Zhu \textsuperscript{a,*}, Xianping Guo \textsuperscript{b}

\textsuperscript{a} Department of Mathematics, South China Normal University, Guangzhou 510631, PR China
\textsuperscript{b} The School of Mathematics and Computational Science, Zhongshan University, Guangzhou 510275, PR China

Received 11 June 2005
Available online 24 March 2006
Submitted by J. Rosinski

Abstract

This paper deals with discrete-time Markov decision processes with \textit{average sample-path costs} (ASPC) in Borel spaces. The costs may have \textit{neither upper nor lower bounds}. We propose new conditions for the existence of \(\varepsilon\)-ASPC-optimal (deterministic) stationary policies in the class of all randomized history-dependent policies. Our conditions are \textit{weaker} than those in the previous literature. Moreover, some sufficient conditions for the existence of ASPC optimal stationary policies are imposed on the \textit{primitive data} of the model. In particular, the \textit{stochastic monotonicity condition} in this paper has first been used to study the ASPC criterion. Also, the approach provided here is slightly different from the “optimality equation approach” widely used in the previous literature. On the other hand, under mild assumptions we show that average expected cost optimality and ASPC-optimality are equivalent. Finally, we use a controlled queueing system to illustrate our results.

\(\copyright\) 2006 Elsevier Inc. All rights reserved.

\textit{Keywords:} Discrete-time Markov decision process; Average sample-path cost; Optimality inequality; Optimal stationary policy

1. Introduction

This paper studies discrete-time Markov decision processes (MDPs) with \textit{average sample-path costs} (ASPC) in Borel spaces. The costs may have \textit{neither upper nor lower bounds} instead of

\(\dagger\) Research partially supported by NCET, RFDP and NSFC.
\(\ast\) Corresponding author.
\textit{E-mail addresses:} zqx22@126.com (Q. Zhu), mcsgxp@mail.sysu.edu.cn (X. Guo).

0022-247X/S – see front matter \(\copyright\) 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmaa.2006.02.050
the case of nonpositive (upper bounds) rewards or nonnegative (lower bounds) costs widely used in MDPs. We propose new conditions for the existence of \( \varepsilon (\geq 0) \)-ASPC-optimal (deterministic) stationary policies in the class of all randomized history-dependent policies.

The long-run average expected cost (AEC) criterion in discrete-time MDPs has been widely studied in the literature; see, for instance, the books [3,5,10–12,14,17,21,22], the survey paper [1] and their extensive references. However, the sample-path costs corresponding to an optimal policy that minimizes the average expected costs may have fluctuations from its expected cost value. To take these fluctuations into account, the ASPC criterion has been proposed and studied; see, for instance, [1–4,7,13,15,16,23,24], as well as the books [11,14] and their extensive bibliographies. To the best of our knowledge, most of the existing works with the ASPC criterion are concentrated on denumerable (or finite) state spaces and/or bounded costs; for instance, see [16] for finite MDPs, and [2–4] for countable MDPs, whereas for MDPs in Borel spaces we know only [1] for bounded costs and [11,13,15,23,24] for unbounded costs. However, all of these works first establish the optimality equation and the existence of a canonical policy \( f^* \), and then show that a canonical policy \( f^* \) is ASPC-optimal. In fact, an ASPC-optimal stationary policy may not be canonical. In this paper, we also deal with the ASPC criterion for MDPs in Borel spaces with possibly unbounded costs. We give another set of optimality conditions, and under which we prove the existence of \( \varepsilon (\geq 0) \)-ASPC-optimal stationary policies that may not be canonical. Our conditions are weaker than those in the previous literature (see [1–4,7,13,15,16,23,24] for instance) because we require that the relative difference of the discounted optimal value function is bounded only in the discount factors \( \alpha \), and remove the irreducibility condition (e.g., Assumption 10.3.5 in [11] and Assumption 3.2(b) in [13]). Moreover, we propose a new condition on the one-step cost which is weaker than the “second order condition” widely used in the previous literature (see [11,13] for instance). On the other hand, to prove the existence of ASPC-optimal stationary policies we also provide a new approach which is slightly different from those used in the previous literature such as [1–4,7,13,15,16,23,24]. More precisely, under our conditions we first provide two optimality inequalities rather than the “optimality equation” in [11,13] for instance, and then not only prove the existence of solutions to the two inequalities but also ensure the existence of ASPC-optimal stationary policies by using the two inequalities. In addition, to verify our assumptions, we further give some sufficient conditions which are imposed on the primitive data of the model. In particular, the “stochastic monotonicity condition” in this paper has first been used to study the ASPC criterion. Also, under mild assumptions we show that AEC-optimality and ASPC-optimality are equivalent. Finally, we apply our results to a controlled queueing system.

The rest of this paper is organized as follows. In Section 2, we introduce the control model and the optimal control problem that we are interested in. After introducing our optimality conditions and some technical preliminaries in Section 3, we study the existence of \( \varepsilon (\geq 0) \)-ASPC-optimal stationary policies in Section 4. Our results are illustrated by a controlled queueing system in Section 5. We conclude in Section 6 with some general remarks.

2. The optimal control problem

**Notation.** If \( X \) is a Borel space (that is, a Borel subset of a complete and separable metric space), we denote by \( B(X) \) the Borel \( \sigma \)-algebra.

In this section we first introduce the control model

\[
\{ S, \{ A(x) : A \in A, x \in S \}, Q(\cdot | x, a), c(x, a) \}, \tag{2.1}
\]
where $S$ and $A$ are the state and the action spaces respectively, which are assumed to be Borel spaces, and $A(x)$ denotes the set of available actions at state $x \in S$. We suppose that the set

$$K := \{(x, a): x \in S, a \in A(x)\}$$

is a Borel subset of $S \times A$. Furthermore,

$$Q(\cdot \mid x, a) \text{ with } (x, a) \in K, \text{ the transition law, is a stochastic kernel on } S \text{ given } K.$$ 

Finally, $c(x, a)$, the cost function, is assumed to be a real-valued measurable function on $K$. (As $c(x, a)$ is allowed to take positive and negative values, it can also be interpreted as a reward function rather than a “cost.”)

To introduce the optimal control problem that we are concerned with, we need to introduce the classes of admissible control policies.

For each $t \geq 0$, let $H_t$ be the family of admissible histories up to time $t$, that is, $H_0 := S$, and $H_t := K \times H_{t-1}$ for each $t \geq 1$.

**Definition 2.1.** A randomized history-dependent policy is a sequence $\pi := (\pi_t, t \geq 0)$ of stochastic kernels $\pi_t$ on $A$ given $H_t$ satisfying

$$\pi_t(A(x) \mid h_t) = 1 \quad \forall h_t = (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x) \in H_t \text{ and } t \geq 0.$$

The class of all randomized history-dependent policies is denoted by $\Pi$. A randomized history-dependent policy $\pi := (\pi_t, t \geq 0) \in \Pi$ is called (deterministic) stationary if there exists a measurable function $f$ from $S$ to $A$ with $f(x) \in A(x)$ for all $x \in S$, such that

$$\pi_t(\{f(x)\} \mid h_t) = \pi_t(\{f(x)\} \mid x) = 1 \quad \forall h_t = (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x) \in H_t \text{ and } t \geq 0.$$

For simplicity, we denote this policy by $f$. The class of all stationary policies is denoted by $F$, which means that $F$ is the set of all measurable functions $f$ from $S$ to $A$ with $f(x) \in A(x)$ for all $x \in S$.

For each $x \in S$ and $\pi \in \Pi$, by the well-known Tulcea’s theorem [5,10,11], there exist a unique probability measure space $(\Omega, \mathcal{F}, \mathbb{P}_x^\pi)$ and a stochastic process $\{x_t, a_t, t \geq 0\}$ defined on $\Omega$ such that, for each $D \in B(S)$ and $t \geq 0$,

$$\mathbb{P}_x^\pi(x_{t+1} \in D \mid h_t, a_t) = Q(D \mid x_t, a_t),$$

(2.3)

with $h_t = (x_0, a_0, \ldots, x_{t-1}, a_{t-1}, x_t) \in H_t$, where $x_t$ and $a_t$ denote the state and action variables at time $t \geq 0$, respectively. The expectation operator with respected to $P_x^\pi$ is denoted by $E_x^\pi$. In particular, when the policy $\pi := f$ is in $F$, the corresponding process $\{x_t\}$ is a time-homogeneous Markov chain with values in $S$ and transition law $Q(\cdot \mid x, f(x))$.

Now we define the ASPC criterion $V_{sp}(\cdot, \cdot)$ as follows: For each $\pi \in \Pi$ and $x \in S$,

$$V_{sp}(x, \pi) := \limsup_{n \to \infty} \frac{\sum_{i=0}^{n-1} c(x_i, a_i)}{n},$$

(2.4)

where the subscript “sp” stands for “sample-path.” Note that $V_{sp}(x, \pi)$ has been defined by the so-called sample-path-costs, $c(x_i, a_i)$; therefore, it is a variable, instead of a number as the AEC criterion defined as
\[ \bar{V}(x, \pi) := \lim sup_{n \to \infty} \frac{E^\pi_x \left[ \sum_{t=0}^{n-1} c(x_t, a_t) \right]}{n}, \]  

see [11,12,17,21,22,25] for instance. Thus, the definition of optimal policies for the ASPC criterion below, is different from that for the AEC criterion.

**Definition 2.2.** For a given \( \epsilon \geq 0 \), a policy \( \pi^* \in \Pi \) is said to be \( \epsilon \)-ASPC-optimal if there exists a constant \( g^* \) such that

\[ P_{\pi^*} \left( V^{sp}(x, \pi^*) \leq g^* + \epsilon \right) = 1 \quad \text{and} \quad P_{\pi} \left( V^{sp}(x, \pi) \geq g^* \right) = 1 \quad \forall x \in S \text{ and } \pi \in \Pi. \]

A 0-ASPC-optimal policy is simply called an ASPC-optimal policy.

The main goal of this paper is to give new conditions for the existence of stationary policies that are ASPC-optimal.

### 3. Optimality conditions

In this section we state conditions for the existence of ASPC-optimal stationary policies and give some preliminary lemmas that are needed to prove our main results.

We shall first introduce two sets of hypotheses. The first one, Assumption A, is a combination of a “Lyapunov-like inequality” condition together with a new condition on the one-step cost \( c \).

**Assumption A.**

1. There exist constants \( b \geq 0 \) and \( 0 < \beta < 1 \) and a (measurable) function \( w \geq 1 \) on \( S \) such that

\[ \int_S w(y) Q(dy \mid x, a) \leq \beta w(x) + b \quad \forall (x, a) \in K. \]  

2. There exist constants \( M > 0 \) and \( 1 < \gamma \leq 2 \) such that \( |c(x, a)|^\gamma \leq M w(x) \) for all \( (x, a) \in K \), with \( w \) as in (3.1).

**Remark 3.1.** Assumption A(1) is well known as a “Lyapunov-like inequality condition,” see [11, p. 121] for instance. Assumption A(2) is new and it is weaker than Assumption 11.3.4 in [11] and Assumption 3.6 in [13]. Obviously, Assumption 11.3.4 in [11] or Assumption 3.6 in [13] implies Assumption A(2) with \( \gamma := 2 \). Moreover, Assumption A(2) implies Assumption A(2) in [8]. In fact, let \( w^*(x) := w(x)^{1/\gamma} \) for all \( x \in S \). Since \( w \geq 1 \), Assumption A(2) above yields

\[ |c(x, a)| \leq M' w^*(x) \leq M' w(x) \quad \text{with } M' := M^{1/\gamma} \quad \forall (x, a) \in K. \]  

The second set of hypotheses we need is the following standard continuity-compactness conditions; see, for instance, [8,11,13,17,22,25] and their references.

**Assumption B.**

1. For each \( x \in S \), \( A(x) \) is compact.
2. For each fixed \( x \in S \), \( c(x, a) \) is lower semicontinuous in \( a \in A(x) \), and the function \( \int_S u(y) Q(dy \mid x, a) \) is continuous in \( a \in A(x) \) for each bounded measurable function \( u \) on \( S \), and also for \( u := w \) as in Assumption A.
To state our optimality conditions, we require some results about the discounted cost criterion. To do this, we use the following notation:

For each fixed discount factor \( 0 < \alpha < 1 \), \( x \in S \) and \( \pi \in \Pi \), the discounted cost \( V_\alpha(x, \pi) \) and the corresponding discounted optimal value function \( V_\alpha^*(x) \) are defined by

\[
V_\alpha(x, \pi) := \mathbb{E}^\pi_x \left[ \sum_{t=0}^{\infty} \alpha^t c(x_t, a_t) \right] \quad \text{and} \quad V_\alpha^*(x) := \inf_{\pi \in \Pi} V_\alpha(x, \pi),
\]

respectively.

To prove the existence of ASPC-optimal stationary policies, in addition to Assumptions A and B, we give another new condition (Assumption C below). To state this assumption, we introduce the following notation:

For the function \( w \geq 1 \) in Assumption A, we define the weighted supremum norm \( \|u\|_w \) for real-valued functions \( u \) on \( S \) by

\[
\|u\|_w := \sup_{x \in S} \left[ w(x)^{-1} |u(x)| \right],
\]

and the Banach space \( B_w(S) := \{u: \|u\|_w < \infty\} \).

**Assumption C.** There exist two functions \( v_1, v_2 \in B_{w^*}(S) \) and some state \( x_0 \in S \) such that

\[
v_1(x) \leq h_\alpha(x) \leq v_2(x) \quad \forall x \in S \text{ and } \alpha \in (0, 1),
\]

where \( h_\alpha(x) := V_\alpha^*(x) - V_\alpha^*(x_0) \) is a so-called relative difference of the function \( V_\alpha^*(x) \), and the function \( w^* \) is as in Remark 3.1.

**Remark 3.2.**

(a) Assumption C is slightly weaker than Assumption C in [8] because the two functions \( v_1, v_2 \) in [8] are required to belong to \( B_w(S) \) being a subset of \( B_{w^*}(S) \), and it is a generalization of (SEN2) in [22, p. 132] and of Assumption 5.4.1(b) in [12] since the difference \( h_\alpha(x) \) is assumed to be bounded below in [12,22].

(b) Assumption C is weaker than Assumptions 3.2 and 3.4 in [13], and Assumption 11.3.4 in [11] because an additional irreducibility condition is required in [11,13].

(c) Assumption C holds if Assumptions A and B as well as the following condition (i.e., Assumption 10.2.2 in [11]) are satisfied: for each \( f \in F \), the Markov process \( \{x_t\} \) is uniform \( w^* \)-exponentially ergodic; that is, there exists a probability measure \( \mu_f \) such that

\[
\left\| Q_f^t (\cdot | x) - \mu_f (\cdot) \right\|_{w^*} \leq R \rho^t w^*(x) \quad \forall x \in S \text{ and } t \geq 0,
\]

where \( R > 0 \) and \( 0 < \rho < 1 \) are constants independent of \( f \), and

\[
\left\| Q_f^t (\cdot | x) - \mu_f (\cdot) \right\|_{w^*} := \sup_{|u| \leq w^*} \mathbb{E}^f_x \left[ u(x_t) - \mu_f (u) \right],
\]

with \( \mu_f (u) := \int_S u(y) \mu_f (dy) \).
(d) Propositions 10.2.4 and 10.2.5 in [11] give some sufficient conditions as well as examples to verify (3.3). These conditions are generalizations of ergodic conditions in [10] and of the minorant condition in [5].

Here, to verify Assumption C we also use some sufficient conditions from [8] which are different from those in [5, 10, 11, 13]. For completeness and ease of reference, we state them as Lemma 3.1 below.

**Lemma 3.1.** Under Assumptions A and B, each one of the following conditions (a) and (b) implies Assumption C:

(a) For each \( f \in F \), there exists a probability measure \( \mu_f \) such that
\[
\sum_{t=0}^{\infty} r^t \left\| Q_f^t(\cdot | x) - \mu_f(\cdot) \right\|_{w^*} \leq Rw^*(x) + L \quad \forall x \in S,
\]
where \( R > 0 \), \( r > 1 \) and \( L \geq 0 \) are constants independent of \( f \).

(b) \( S = [0, \infty) \) and the two following conditions are satisfied:

(b1) the process \( \{x_t\} \) is stochastically ordered (monotonic) for each \( f \in F \);  
(b2) the function \( w \) in Assumption A is nondecreasing and such that
\[
\int_S w(y) Q(dy | x, f(x)) \leq \beta w(x) + b I_{\{0\}}(x) \quad \forall f \in F \text{ and } x \geq 0,
\]
where \( I_D \) denotes the indicator function of any set \( D \).

**Proof.** Obviously, part (a) follows from the proof of Lemma 3.3(a) in [8]. We now prove (b). In fact, by Jensen’s inequality and (b2) we have
\[
\int_S w^*(y) Q(dy | x, f(x)) \leq \beta' w^*(x) + b' I_{\{0\}}(x) \quad \forall f \in F \text{ and } x \geq 0,
\]
where \( \beta' := \beta^{1/\gamma} \) and \( b' := b^{1/\gamma} \), and \( \gamma \) is as in Assumption A. From this inequality and (b1) as well as the proof of (14) in [20] we have that for each \( x \in S, f \in F \) and \( 1 < r < \frac{1}{\beta'} \),
\[
\sum_{t=0}^{\infty} r^t \left\| Q_f^t(\cdot | x) - \mu_f(\cdot) \right\|_{w^*} \leq \frac{2}{1 - \beta' r} \left[ w^*(x) + \frac{b'}{1 - \beta'} \right],
\]
which together with (a), gives Assumption C. □

**Remark 3.3.** From the proof of Lemma 3.1, we conclude that Lemma 3.1(b) implies Lemma 3.1(a). Obviously, the conditions in Lemma 3.1 are different from those in [1, 5, 6, 10–12]. In particular, the stochastic monotonicity condition (b1) has been used to verify Assumption C.

Under Assumptions A–C, we can obtain several lemmas, which are needed to prove our main results.

**Lemma 3.2.** Under Assumptions A–C, the following statements hold.

There exist a unique constant \( g^* \), two functions \( h_k^* \in B_{w^*}(S) \) (\( k = 1, 2 \), with \( w^* \) as in Remark 3.1), and a stationary policy \( f^* \in F \) satisfying the two average cost optimality inequalities
Proof. By Jensen’s inequality and Assumption A(1) we have
\[ \int_S w^*(y) Q(dy|x,a) \leq \beta' w^*(x) \quad \forall (x,a) \in K, \] (3.5)
where \( \beta' := \beta^{1/\gamma} \) and \( b' := b^{1/\gamma} \), and \( \gamma \) is as in Assumption A. This inequality together with (3.2) gives Assumption A in [8]. Thus, under Assumptions A–C, using (3.2) and (3.5) we see that all the conditions in [8] also hold after the function \( w \) in [8] is replaced by \( w^* \) here. Therefore, from Theorem 4.1 in [8] we conclude that Lemma 3.2 holds. \( \square \)

**Lemma 3.3.** Under the hypotheses of Lemma 3.1(b), Assumptions A and B, the following statements hold.

For each \( f \in F \), the Markov process \( \{x_t\} \) is uniform \( w^* \)-exponentially ergodic; that is, there exists a probability measure \( \mu_f \) such that
\[ \| Q^t_f (\cdot | x) - \mu_f (\cdot) \|_{w^*} \leq R \rho^t w^*(x) \quad \forall x \in S \text{ and } t \geq 0, \] (3.6)
where \( R > 0 \) and \( 0 < \rho < 1 \) are constants independent of \( f \), and the function \( w^* \) is as in Remark 3.1.

**Proof.** As in the proof of Lemma 3.1(b), we have that for each \( f \in F, x \in S \) and \( 1 < r < \frac{1}{b'} \),
\[ \sum_{t=0}^{\infty} r^t \| Q^t_f (\cdot | x) - \mu_f (\cdot) \|_{w^*} \leq \frac{2}{1 - b'} \left[ w^*(x) + \frac{b'}{1 - b'} \right], \] (3.7)
where \( \beta' := \beta^{1/\gamma} \) and \( b' := b^{1/\gamma} \), and \( \gamma \) is as in Assumption A. Thus, from (3.7) we see that for each \( f \in F \), there exist positive constants (independent of \( f \)) \( R_0 \) and \( \rho_0 \) with \( 0 < \rho_0 \leq r^{-1} \) such that
\[ \| Q^t_f (\cdot | x) - \mu_f (\cdot) \|_{w^*} \leq R_0 \rho_0^t w^*(x) \quad \forall x \in S, \]
which gives Lemma 3.3 with \( R := R_0 \) and \( \rho := \rho_0 \). \( \square \)

**Lemma 3.4.** Suppose that the hypotheses of Lemma 3.3 hold and let \( f \in F \) be any stationary policy. Then for each \( x \in S \), we have

\( a) \quad \tilde{V}(x, f) = g(f), \)

where \( g(f) := \int_S c(y, f(y)) \mu_f (dy). \)

\( b) \quad V_{sp}(x, f) = g(f) \quad P^f_x - \text{a.s.} \) (3.9)
Proof. (a) Combining (3.2) and (3.6) we have that for each $f \in F$ and $x \in S$,
\[
\left| E_x^t \left[ c(x_t, f(x_t)) \right] - g(f) \right| \leq M' R \rho^t w^*(x), \quad t = 0, 1, \ldots,
\]
and so
\[
\left| \sum_{t=0}^{n-1} E_x^t \left[ c(x_t, f(x_t)) \right] - ng(f) \right| \leq M' R w^*(x) \frac{1 - \rho^n}{1 - \rho} \leq M' R w^*(x) \frac{1}{1 - \rho} \quad \forall n \geq 1.
\]
Multiply both sides of (3.11) by $\frac{1}{n}$ and then take $\limsup$ as $n \to \infty$ to obtain (3.8).

(b) By the strong law of large numbers (see Theorem 11.2.1(a) in [11] for instance) and the definition (2.4) of the long-run ASPC, we obtain (3.9). \qed

4. Existence of ASPC-optimal stationary policies

In this section, we first use Assumptions A–C to prove the existence of an ASPC-optimal stationary policy. Then, under mild assumptions we show that AEC-optimality and ASPC-optimality are equivalent.

Theorem 4.1. Under Assumptions A–C, the following statements hold:

(a) There exist a unique constant $g^*$, two functions $h_k^* \in B_{w^*}(S)$ ($k = 1, 2$, with $w^*$ as in Remark 3.1), and a stationary policy $f^* \in F$ satisfying the two average cost optimality inequalities
\[
g^* + h_1^*(x) \leq \min_{a \in \Lambda(x)} \left\{ c(x, a) + \int_S h_1^*(y) Q(dy \mid x, a) \right\} \quad \forall x \in S; \quad (4.1)
\]
\[
g^* + h_2^*(x) \geq \min_{a \in \Lambda(x)} \left\{ c(x, a) + \int_S h_2^*(y) Q(dy \mid x, a) \right\} \quad (4.2)
\]
\[
= c(x, f^*(x)) + \int_S h_2^*(y) Q(dy \mid x, f^*(x)) \quad \forall x \in S. \quad (4.3)
\]

(b) The policy $f^*$ in (a) is ASPC-optimal, and
\[
P_x^{f^*} \left( V_{sp}(x, f^*) = g^* \right) = 1 \quad \text{for all } x \in S.
\]

(c) Any stationary policy $f \in F$ is ASPC-optimal if it realizes the minimum of (4.2).

(d) For a given $\varepsilon \geq 0$ and $f \in F$, if there exists a function $h \in B_{w^*}(S)$ such that
\[
g^* + h(x) \geq c(x, f(x)) + \int_S h(y) Q(dy \mid x, f(x)) - \varepsilon \quad \forall x \in S,
\]
then $f$ is $\varepsilon$-ASPC-optimal.
Proof. (a) By Lemma 3.2 we see that part (a) is true.
(b) To prove (b), for each $x \in S$ and $\pi = (\pi_t, t \geq 0) \in P_I$, let

$$F_n := \sigma \{ x_0, a_0, x_1, a_1, \ldots, x_n, a_n \} \quad \forall n \geq 0,$$

$$L_n(x, \pi, h_k^n) := \sum_{t=1}^{n} Z_t(x, \pi, h_k^n) \quad \forall n \geq 1 \text{ and } k = 1, 2. \quad (4.5)$$

where

$$Z_t(x, \pi, h_k^n) := h_k^n(x_t) - E_{x}^{\pi} \left[ h_k^n(x_t) \mid F_{t-1} \right] \quad \forall t \geq 1 \text{ and } k = 1, 2. \quad (4.6)$$

We will show that the sequences $L_n(x, \pi, h_k^n) (k = 1, 2)$ are $P_x^{\pi}$-martingales with respect to \{\mathcal{F}_n\}, that is

$$E_{x}^{\pi} \left[ L_n(x, \pi, h_k^n) \mid \mathcal{F}_{n-1} \right] = L_{n-1}(x, \pi, h_k^n) \quad P_{x}^{\pi}\text{-a.s. } \forall n \geq 1 \text{ and } k = 1, 2.$$

In fact, it follows from (4.6) that

$$\left| Z_t(x, \pi, h_k^n) \right| \leq \left| h_k^n(x_t) \right| + E_{x}^{\pi} \left[ \left| h_k^n(x_t) \right| \mid F_{t-1} \right]$$

$$
\leq \| h_k^n \|_{w^*} \left\{ w^*(x_t) + E_{x}^{\pi} \left[ w^*(x_t) \mid F_{t-1} \right] \right\}, \quad \text{for } k = 1, 2. \quad (4.7)
$$

and so

$$E_{x}^{\pi} \left[ \left| Z_t(x, \pi, h_k^n) \right| \right] \leq 2 \| h_k^n \|_{w^*} E_{x}^{\pi} \left[ w^*(x_t) \right] \quad \text{for } k = 1, 2. \quad (4.8)$$

Moreover, from the proof of Lemma 3.1 in [8] and (3.5) we have

$$E_{x}^{\pi} \left[ w^*(x_t) \right] \leq \beta^{t} w(x) + \frac{b}{1 - \beta},$$

which together with (4.8) shows that $L_n(x, \pi, h_k^n) (k = 1, 2)$ are $P_x^{\pi}$-integrable for each $n$.

Obviously, combining (4.5) and (4.6) we conclude that $L_n(x, \pi, h_k^n) (k = 1, 2)$ are $\mathcal{F}_n$-measurable and that

$$E_{x}^{\pi} \left[ L_n(x, \pi, h_k^n) \mid \mathcal{F}_{n-1} \right] = L_{n-1}(x, \pi, h_k^n) \quad P_{x}^{\pi}\text{-a.s. } \forall n \geq 1 \text{ and } k = 1, 2.$$

Therefore, $L_n(x, \pi, h_k^n) (k = 1, 2)$ are $P_x^{\pi}$-martingales.

On the other hand, by Assumption A and Lemma 3.1 in [8] we have

$$E_{x}^{\pi} \left[ w(x_t) \right] \leq \beta^{t} w(x) + \frac{b}{1 - \beta} \quad \forall t \geq 0, \quad (4.9)$$

which together with (4.5)–(4.7) and Jensen’s inequality, gives

$$E_{x}^{\pi} \left[ L_{n+1}(x, \pi, h_k^n) - L_n(x, \pi, h_k^n) \right]^{\gamma}$$

$$= E_{x}^{\pi} \left[ Z_{n+1}(x, \pi, h_k^n) \right]^{\gamma} \quad \text{(by (4.5))}$$

$$\leq E_{x}^{\pi} \left\{ \left| h_k^n \right|^{\gamma}_{w^*} \left[ w^*(x_{n+1}) + E_{x}^{\pi} \left[ w^*(x_{n+1}) \mid \mathcal{F}_n \right] \right]^{\gamma} \right\} \quad \text{(by (4.7))}$$

$$\leq 2^{\gamma} \| h_k^n \|^{\gamma}_{w^*} E_{x}^{\pi} \left[ w(x_{n+1}) + E_{x}^{\pi} \left[ w(x_{n+1}) \mid \mathcal{F}_n \right] \right] \right\}$$

$$\leq 4 \| h_k^n \|^{\gamma}_{w^*} E_{x}^{\pi} \left[ w(x_{n+1}) + E_{x}^{\pi} \left[ w(x_{n+1}) \mid \mathcal{F}_n \right] \right] \right\}$$

$$\leq 8 \| h_k^n \|^{\gamma}_{w^*} E_{x}^{\pi} \left[ w(x_{n+1}) \right] \right\} \quad \text{(by Jensen’s inequality)}$$

$$\leq 8 \| h_k^n \|^{\gamma}_{w^*} \left[ \beta^{n+1} w(x) + \frac{b}{1 - \beta} \right] \quad \text{for } k = 1, 2. \quad (4.10)$$
The second inequality in (4.10) follows from the elementary inequality:
\[(a + b)^\gamma \leq 2^\gamma (a^\gamma + b^\gamma) \quad \forall a > 0, \ b > 0 \text{ and } 1 < \gamma \leq 2.\]

From (4.10) we see that \(E_\pi^\gamma \left[ L_{n+1}^\gamma (x, \pi, h_k^*) - L_n(x, \pi, h_k^*) \right] (k = 1, 2)\) are bounded in \(n \geq 1\). Thus, by the martingale stability theorem (see Theorem 2.18 in [9] for instance), we have
\[
\lim_{n \to \infty} \frac{1}{n} L_n(x, \pi, h_k^*) = 0 \quad P_\pi^\pi \text{-a.s. (}k = 1, 2). \tag{4.11}
\]
Moreover, from (4.9) it follows that
\[
E_\pi^\pi \left[ \sum_{t=1}^{\infty} t^{-\gamma} w(x_t) \right] < \infty, \quad P_\pi^\pi \text{-a.s.,}
\]
and so
\[
\sum_{t=1}^{\infty} t^{-\gamma} w(x_t) < \infty, \quad P_\pi^\pi \text{-a.s.,}
\]
which implies
\[
\lim_{t \to \infty} t^{-\gamma} w(x_t) = 0 \quad P_\pi^\pi \text{-a.s.} \tag{4.12}
\]
Since \(w^*(x) = w(x)^{1/\gamma}\) for all \(x \in S\), (4.12) yields
\[
\lim_{t \to \infty} \frac{w^*(x_t)}{t} = 0 \quad P_\pi^\pi \text{-a.s.} \tag{4.13}
\]
Also, from part (a) it follows that \(h_k^* \in B_{w^*} (S) (k = 1, 2)\), which together with (4.13) gives
\[
\lim_{t \to \infty} \frac{h_k^*(x_t)}{t} = 0 \quad (k = 1, 2). \tag{4.14}
\]
Now we consider the two so-called discrepancy functions \(\Delta_k : K \to R (k = 1, 2)\) defined as
\[
\Delta_1(x, a) := c(x, a) + \int_S h_1^*(y) Q(dy \ | \ x, a) - g^* - h_1^*(x) \quad \forall (x, a) \in K; \tag{4.15}
\]
\[
\Delta_2(x, a) := c(x, a) + \int_S h_2^*(y) Q(dy \ | \ x, a) - g^* - h_2^*(x) \quad \forall (x, a) \in K. \tag{4.16}
\]
Obviously, by (4.1) we have \(\Delta_1(x, a) \geq 0\) for all \((x, a) \in K\). Also, from (2.3) we rewrite \(Z_t(x, \pi, h_k^*) (k = 1, 2)\) as
\[
Z_t(x, \pi, h_1^*) = h_1^*(x_t) - \int_S h_1^*(y) Q(dy \ | \ x_t, a_t) \quad \forall t \geq 1, \quad \text{and}
\tag{4.17}
\]
\[
Z_t(x, \pi, h_2^*) = h_2^*(x_t) - \int_S h_2^*(y) Q(dy \ | \ x_t, a_t) \quad \forall t \geq 1, \tag{4.18}
\]
respectively. Then, combining (4.15) and (4.17) we see that (4.5) becomes
\[
L_n(x, \pi, h_1^*) = h_1^*(x_n) - h_1^*(x) - \sum_{t=0}^{n-1} \Delta_1(x_t, a_t) + \sum_{t=0}^{n-1} c(x_t, a_t) - n g^* \quad \forall n \geq 1,
\]
and so
\[ \sum_{t=0}^{n-1} c(x_t, a_t) \geq ng^* + L_n(x, \pi, h_1^*) - h_1^*(x_n) + h_1^*(x) \quad \forall n \geq 1. \tag{4.19} \]

Multiply both sides of (4.19) by \( \frac{1}{n} \) and then take liminf as \( n \to \infty \) to obtain (by (4.11) and (4.14))
\[ V_{sp}(x, \pi) \geq \lim_{n \to \infty} \inf \sum_{t=0}^{n-1} \frac{c(x_t, a_t)}{n} \geq g^* \quad P_x^\pi \text{-a.s.} \forall \pi \in \Pi \text{ and } x \in S. \tag{4.20} \]

On the other hand, from (4.3) it follows that \( \Delta_2(x, f^*(x)) \leq 0 \) for all \( x \in S \).
Similarly, by (4.16), (4.18) and (4.5) we have
\[ L_n(x, f^*, h_2^*) = h_2^*(x_n) - h_2^*(x) - \sum_{t=0}^{n-1} \Delta_2(x_t, f^*(x_t)) \]
\[ + \sum_{t=0}^{n-1} c(x_t, f^*(x_t)) - ng^* \quad \forall x \in S \text{ and } n \geq 1, \]
and so
\[ \sum_{t=0}^{n-1} c(x_t, f^*(x_t)) \leq ng^* + L_n(x, f^*, h_2^*) - h_2^*(x_n) + h_2^*(x) \quad \forall x \in S \text{ and } n \geq 1. \tag{4.21} \]

Thus, as in the proof of (4.20), from (4.21) it follows that
\[ V_{sp}(x, f^*) \leq g^* \quad P_x^{f^*} \text{-a.s.} \forall x \in S. \tag{4.22} \]

Therefore, combining (4.20) and (4.22) we see that for each \( x \in S \) and \( \pi \in \Pi \),
\[ P_x^{f^*}(V_{sp}(x, f^*) = g^*) = 1 \quad \text{and} \quad P_x^\pi(V_{sp}(x, \pi) \geq g^*) = 1, \tag{4.23} \]
which implies that \( f^* \) is ASPC-optimal.
(c) Obviously, (c) follows from part (b).
(d) Let
\[ \Delta_h(x, f(x)) := c(x, f(x)) + \int_S h(y) Q(dy \mid x, f(x)) - g^* - h(x). \]
Then, from (4.4) it follows that
\[ \Delta_h(x, f(x)) \leq \epsilon \quad \text{for all } x \in S. \]
Thus, as in the proof of (4.22), we have
\[ P_x^f(V_{sp}(x, f) \leq g^* + \epsilon) = 1 \quad \forall x \in S, \]
which together with (4.23), gives part (d). \( \square \)
Remark 4.1.

(a) Theorem 4.1 is our first main result: part (a) establishes the two optimality inequalities (4.1) and (4.2), and the existence of an AEC-optimal stationary policy \( f^* \), whereas part (b) further shows that the AEC-optimal stationary policy \( f^* \) is ASPC-optimal. It should be noted that the approach provided in this paper is different from those used in the previous literature. More explicitly, under our conditions we first provide two optimality inequalities rather than the “optimality equation” in the previous literature such as [11,13], and then not only prove the existence of solutions to the two inequalities but also ensure the existence of ASPC-optimal stationary policies by using the two inequalities.

(b) Our conditions are weaker than those in the previous literature because the function \( h_{\alpha}(x) \) may have neither upper nor lower bounds and the irreducibility condition (e.g., Assumption 10.3.5 in [11] and Assumption 3.2(b) in [13]) is removed. Moreover, we provide some new sufficient conditions which are imposed on the model itself and are easily verified.

Theorem 4.1 proves the AEC-optimal stationary policy that realizes the minimum of (4.2) is ASPC-optimal. It is natural to question if an ASPC-optimal stationary policy must be AEC-optimal? Under mild assumptions, the following theorem gives an affirmative answer.

**Theorem 4.2.** Under the hypotheses of Lemma 3.1(b), Assumptions A and B, the following statements hold:

(a) All the conclusions in Theorem 4.1 hold.
(b) Any stationary policy \( \bar{f} \in F \) is AEC-optimal if it is ASPC-optimal.
(c) Conversely, any stationary policy \( \bar{f} \in F \) is ASPC-optimal if it is AEC-optimal.
(d) Any stationary policy \( \bar{f} \in F \) is AEC-optimal if and only if it is ASPC-optimal.

**Proof.** Obviously, part (a) follows from Lemma 3.3, Theorem 4.1 and Remark 3.3. We now prove (b). In fact, from (4.1) it follows that

\[
g^* + h_1^*(x_t) \leq c(x_t, f(x_t)) + \int_S h_1^*(y)Q(dy | x_t, f(x_t)) \quad \forall f \in F \text{ and } t \geq 0,
\]

which together with (2.3), gives

\[
g^* + E_x^f [h_1^*(x_t)] \leq E_x^f [c(x_t, f(x_t))] + E_x^f [h_1^*(x_{t+1})] \quad \forall t \geq 0,
\]

and so

\[
g^* + \frac{h_1^*(x)}{N} \leq \frac{E_x^f [\sum_{t=0}^{N-1} c(x_t, f(x_t))] + E_x^f [h_1^*(x_N)]}{N} \quad \forall N \geq 1.
\]

From the proof of Theorem 4.1 in [8], (4.24) yields

\[
g^* \leq \bar{V}(x, f) \quad \forall f \in F,
\]

and so

\[
g^* \leq \inf_{f \in F} \bar{V}(x, f).
\]

On the other hand, since \( \bar{f} \) is ASPC-optimal, from Lemma 3.4 it follows that
\[ g^* = V_{sp}(x, \tilde{f}) = g(\tilde{f}) = \bar{V}(x, \tilde{f}), \]

which together with (4.25) yields

\[ \bar{V}(x, \tilde{f}) \leq \inf_{f \in F} \bar{V}(x, f). \]

Thus, \( \tilde{f} \) is AEC-optimal.

(c) As in the proof of Theorem 4.1(b), we see that (4.20) is true. That is,

\[ V_{sp}(x, \pi) \geq g^* \quad P_x^\pi \text{-a.s. } \forall \pi \in \Pi \text{ and } x \in S. \quad (4.26) \]

On the other hand, since \( \tilde{f} \) is AEC-optimal, then Lemma 3.4 yields

\[ g^* = \bar{V}(x, \tilde{f}) = g(\tilde{f}) = V_{sp}(x, \tilde{f}) \quad \forall x \in S. \quad (4.27) \]

Combining (4.26) and (4.27), we have

\[ P_x(\bar{V}_{sp}(x, \tilde{f}) = g^*) = 1 \quad \text{and} \quad P_x(\bar{V}_{sp}(x, \pi) \geq g^*) = 1 \quad \forall x \in S \text{ and } \pi \in \Pi. \]

Hence, \( \tilde{f} \) is ASPC-optimal.

(d) Obviously, (d) follows from parts (b) and (c). \( \square \)

Remark 4.2. Theorem 4.2 is our second main result, and under mild assumptions it shows that AEC-optimality and ASPC-optimality are equivalent.

5. An example

In this section we apply our results to a controlled queueing system.

Example 5.1 (A controlled queueing system). Consider a controlled queueing system in which the state variable denotes a number of customers in the system. The arrival rate \( \lambda \) and the service rate \( \mu \) are assumed to be controlled by a decision-maker. Here we regard the vector \( a := (\lambda, \mu) \) as the control action. When the system’s state is at \( x \in S := \{0, 1, \ldots\} \), the decision maker takes an action \( a \) from a given set \( A(x) \equiv [\lambda_1, \lambda_2] \times [\mu_1, \mu_2] \) with \( \lambda_2 > \lambda_1 > 0 \) and \( \mu_2 > \mu_1 > 0 \), which increases or decreases the arrival rates and the service rates given by (5.1)–(5.3) below. The action incurs a cost \( \bar{c}(x, a) \). In addition, the decision-maker obtains a reward \( px \) during the period which the system remains in state \( x \), where the unit reward caused by a customer is presented by the constant \( p > 0 \).

We now formulate this system as a discrete-time Markov decision process. The corresponding transition probability \( Q(y \mid x, a) \) and cost rates \( c(x, a) \) are given as follows:

\[
\begin{align*}
Q(0 \mid 0, a) & := 1 - \frac{\lambda}{\lambda_2 + \mu_2}, \\
Q(1 \mid 0, a) & := \frac{\lambda}{\lambda_2 + \mu_2} \quad \forall a \in [\lambda_1, \lambda_2] \times [\mu_1, \mu_2]; \\
Q(0 \mid 1, a) & := \frac{\mu_2}{\lambda_2 + \mu_2}, \\
Q(1 \mid 1, a) & := \frac{\lambda_2 - \lambda}{\lambda_2 + \mu_2}, \\
Q(2 \mid 1, a) & := \frac{\lambda}{\lambda_2 + \mu_2} \quad \forall a \in [\lambda_1, \lambda_2] \times [\mu_1, \mu_2].
\end{align*}
\]  

(5.1)
For each \( x \geq 2 \) and \( a \in [\lambda_1, \lambda_2] \times [\mu_1, \mu_2] \),

\[
Q(y \mid x, a) :=
\begin{cases}
\frac{\mu}{\lambda_2 + \mu_2}, & \text{if } y = x - 2; \\
\frac{\mu_2 - \mu}{\lambda_2 + \mu_2}, & \text{if } y = x - 1; \\
\frac{\lambda_2 - \lambda}{\lambda_2 + \mu_2}, & \text{if } y = x; \\
\frac{\lambda}{\lambda_2 + \mu_2}, & \text{if } y = x + 1; \\
0, & \text{otherwise};
\end{cases}
\]

(5.3)

\[
c(x, a) := \bar{c}(x, a) - px \quad \text{for } (x, a) \in K := \{(x, a): x \in S, a \in A(x)\}.
\]

(5.4)

We aim to find conditions that ensure the existence of an ASPC-optimal stationary policy. To do this, we consider the following assumptions:

(E1) The parameter \( \lambda \) satisfies that \( \mu_2 > \lambda_2 e \), where \( e \) is the well-known exponential constant.

(E2) The function \( \bar{c}(x, a) \) is continuous in \( a \in A(x) = [\lambda_1, \lambda_2] \times [\mu_1, \mu_2] \) for each fixed \( x \in S \), and such that \( \bar{c}^*(x) := \sup_{a \in A(x)} |\bar{c}(x, a)| < Cx \) for all \( x \in S \) and some constant \( C > 0 \).

Under these conditions, we have the following.

**Proposition 5.1.** Under Assumptions (E1) and (E2), the above controlled queueing system satisfies Assumptions A–C. Therefore (by Theorem 4.1), there exists an ASPC-optimal stationary policy.

**Proof.** We shall first verify Assumption A. Let \( \rho := \frac{\mu_2 + \lambda_2 e^2}{e(\lambda_2 + \mu_2)} \), \( w(x) := e^x \) for all \( x \in S \). Obviously, by (E1) we obtain \( 0 < \rho < 1 \). Then, combining (5.1) and (5.2) we have

\[
\sum_{y \in S} Q(y \mid 0, a) w(y) = 1 - \frac{\lambda}{\lambda_2 + \mu_2} + \frac{\lambda e}{\lambda_2 + \mu_2} \\
\leq 1 + \frac{(e - 1)\lambda_2}{\lambda_2 + \mu_2} \\
= \rho w(0) + \frac{(e - 1)\mu_2}{e(\lambda_2 + \mu_2)};
\]

(5.5)

\[
\sum_{y \in S} Q(y \mid 1, a) w(y) = \frac{\mu_2}{\lambda_2 + \mu_2} + \frac{\lambda_2 - \lambda}{\lambda_2 + \mu_2} e + \frac{\lambda}{\lambda_2 + \mu_2} e^2 \\
= \frac{\mu_2 + (\lambda_2 - \lambda)e + \lambda e^2}{e(\lambda_2 + \mu_2)} w(1) \\
\leq \rho w(1).
\]

(5.6)

Moreover, for each \( x \geq 2 \) and \( a \in [\lambda_1, \lambda_2] \times [\mu_1, \mu_2] \), from (5.3) it follows that

\[
\sum_{y \in S} Q(y \mid x, a) w(y) = \frac{\mu}{\lambda_2 + \mu_2} e^{x-2} + \frac{\mu_2 - \mu}{\lambda_2 + \mu_2} e^{x-1} + \frac{\lambda_2 - \lambda}{\lambda_2 + \mu_2} e^{x} + \frac{\lambda}{\lambda_2 + \mu_2} e^{x+1} \\
= \frac{\mu e(\mu_2 - \mu) + e^2(\lambda_2 - \lambda) + \lambda^3 e^{x}}{e^2(\lambda_2 + \mu_2)} w(x)
\]
\[ \leq \frac{e\mu_2 + \mu_1(1 - e) + \lambda_2 e^3}{e^2(\lambda_2 + \mu_2)} w(x) \leq \rho w(x). \]  

(5.7)

By the inequalities (5.5)–(5.7) we obtain

\[ \sum_{y \in \mathcal{S}} Q(y \mid x, a)w(y) \leq \rho w(x) + \frac{(e - 1)\mu_2}{e(\lambda_2 + \mu_2)} \mathcal{I}_{\{0\}}(x) \quad \forall a \in A(x) \text{ and } x \in \mathcal{S} \]  

(5.8)

\[ \leq \rho w(x) + \frac{(e - 1)\mu_2}{e(\lambda_2 + \mu_2)} \quad \forall a \in A(x) \text{ and } x \in \mathcal{S}, \]  

(5.9)

which gives Assumption A(1) with \( b := \frac{(e - 1)\mu_2}{e(\lambda_2 + \mu_2)} \) and \( \beta := \rho < 1 \) defined as above.

On the other hand, since \( e^x \geq \frac{x^2}{2} \) and \( 1 < \gamma \leq 2 \), from (5.4) and (E2),

\[ \sup_{a \in A(x)} |c(x, a)|^\gamma \leq (p + C)^2 x^2 \leq 2(p + C)^2 w(x) \quad \text{for all } x \in \mathcal{S}, \]  

which verifies Assumption A(2) with \( M := 2(p + C)^2 \) and \( 1 < \gamma \leq 2 \). And so Assumption A follows.

We now verify Assumption B. Obviously, by (5.1)–(5.4) and the model’s description as well as (E2) we see that Assumption B is satisfied.

Finally, we verify Assumption C. Obviously, from (5.1)–(5.3) we have, for each fixed \( f \in F \):

\[ \sum_{y \geq k} Q(y \mid x, f(x)) \leq \sum_{y \geq k} Q(y \mid x', f(x')) \quad \forall x, x', k \in \mathcal{S} \text{ and } x < x', \]  

which together with Theorem 7.4.1 in [18] implies that, for each \( f \in F \), the corresponding Markov process \( \{x_t\} \) is stochastically monotone. Thus, by (5.8) and Lemma 3.1(b) we see that Assumption C is satisfied. \( \square \)

6. Concluding remarks

In the previous sections we have studied the average sample-path cost (ASPC) optimality problem for discrete-time Markov decision processes in Borel spaces, and the relations between average expected cost (AEC) optimality and ASPC-optimality. It should be noted that the AEC criterion and the ASPC criterion are two different criteria, and some examples in [19] are given to illustrate their difference. In this paper, we propose new conditions for the existence of \( \varepsilon \)-ASPC-optimal (deterministic) stationary policies in the class of all randomized history dependent policies. Moreover, under mild assumptions we also have shown that AEC-optimality and ASPC-optimality are equivalent. It should be mentioned that the approach provided here is different from “optimality equation approach” widely used in the previous literature (see [11,13] for instance).

References