SIAM J. CONTROL OPTIM.
Vol. 44, No. 1, pp. 29–48

OPTIMAL CONTROL OF ERGODIC CONTINUOUS-TIME MARKOV CHAINS WITH AVERAGE SAMPLE-PATH REWARDS*

XIANPING GUO† AND XI-REN CAO‡

Abstract. In this paper we study continuous-time Markov decision processes with the *average sample-path reward* (ASPR) criterion and possibly unbounded transition and reward rates. We propose conditions on the system's *primitive data* for the existence of ϵ-ASPR-optimal (deterministic) stationary policies in a class of randomized Markov policies satisfying some additional continuity assumptions. The proof of this fact is based on the *time discretization* technique, the martingale stability theory, and the concept of potential. We also provide both policy and value iteration algorithms for computing, or at least approximating, the ϵ-ASPR-optimal stationary policies. We illustrate with examples our main results as well as the difference between the ASPR and the average expected reward criteria.

Key words. average sample-path reward, continuous-time Markov chain, optimal stationary policy, policy and value iteration algorithms

AMS subject classifications. 90C40, 93E20

DOI. 10.1137/S0363012903420875

1. Introduction.

Markov decision processes (MDPs) with the long-run *average expected reward* (AER) criterion have been widely studied in literature; see, for instance, the books [1, 6, 12, 22, 23, 25, 31, 32, 33, 35], the survey paper [3], and their extensive references. However, the *sample-path* reward corresponding to an optimal policy that maximizes the average expected rewards may have fluctuations from its expected reward value. To take these fluctuations into account, the *average sample-path reward* (ASPR) criterion has been proposed and studied; see, for instance, [3, 10, 15, 23, 24] and their extensive bibliographies. To the best of our knowledge, all the existing works with the ASPR criterion are on *discrete-time* MDPs. On the other hand, many real-world problems, for instance, in communication engineering, queueing systems, and other control problems, require continuous-time models. Therefore, there is a large amount of works in literature on continuous-time MDPs; see, for instance, [4, 5, 16, 18, 19, 20, 21, 26, 27, 29, 31, 35, 37, 39] and their references. All of these works, however, consider only the AER criterion. Our paper is a first attempt to fill the gap between the works on discrete-time MDPs with the ASPR criterion and those on continuous-time MDPs with the AER criterion.

Denumerable continuous-time MDPs are specified by the system's four *primitive data*: a countable state space S; action sets $A(i)$, which may depend on the current state $i \in S$; transition rates $q(j|i,a)$ with $a \in A(i)$ and $j \in S$; and reward rates $r(i,a)$ with $a \in A(i)$. In this paper, we consider these MDPs with the ASPR criterion in the class of randomized Markov policies satisfying some *additional* continuity assumptions.

*Received by the editors January 7, 2003; accepted for publication (in revised form) August 18, 2004; published electronically June 14, 2005. This work was supported by a grant from Hong Kong UGC.

http://www.siam.org/journals/sicon/44-1/42087.html

†The School of Mathematics and Computational Science, Zhongshan University, Guangzhou 510275, People's Republic of China (mcsgxp@zsu.edu.cn). The research of this author has been supported by the Natural Science Foundations of China and Guangdong Province, by EYTP, NCET, and by Zhongshan University Advanced Research Center, China.

‡Corresponding author. Department of Electrical and Electronic Engineering, The Hong Kong University of Science and Technology, Hong Kong (eecao@ust.hk).

The state processes here are possibly *nonhomogeneous* continuous-time Markov chains with possibly *unbounded* transition rates, and the reward rates may have *neither upper nor lower bounds*. Under suitable conditions on the primitive data, we first prove the existence of a solution to the optimality equation. The proof is constructive, using *policy iteration*, which is based on the concept of potentials [8, 7] and is rather different from both the "vanishing discount approach" in [16, 18, 19, 21, 26] and the "uniformization technique" in [29, 31, 36]. We then establish the existence of $\epsilon(\geq 0)$-ASPR-optimal stationary policies by introducing a *time-discretization approach* to continuous-time martingales and by using the *extended generator* technique. This approach is different from those used for the discrete-time case; see, for instance, [3, 6, 10, 15, 23, 24]. Also, we provide both policy and value iteration algorithms for computing, or at least approximating (when the algorithms take infinitely many steps to converge), $\epsilon(\geq 0)$-ASPR-optimal stationary policies. Furthermore, we use several examples to explain our conditions and to show the difference between the AER and ASPR criteria.

The policy iteration approach developed in this paper to establish a solution to the optimality equation does not require any result about discounted continuous-time MDPs. Thus, this approach is simple and direct. Also, our method to prove the existence of an ASPR-optimal stationary policy is straightforward and different from those in [29, 31, 36, 37], which require the *equivalence* between continuous- and discrete-time MDPs as well as results about discrete-time MDPs. Finally, it should be mentioned that the ergodicity results about continuous-time Markov chains and the convergence results for continuous martingales available in the literature *cannot* be applied to our problems because in this paper the Markov chains may be *nonhomogeneous* and the associated reward and transition rates may be *time-dependent* and *unbounded*. In addition, a key feature of our results is that the conditions are imposed on the *primitive data* (see (2.1)) and can be easily verified.

The rest of this paper is organized as follows. In section 2, we introduce the control model and the optimal control problem considered in this paper. After some technical preliminaries developed in section 3, we study the existence of the $\epsilon(\geq 0)$-ASPR optimal stationary policies in section 4. The policy and value iteration algorithms are described in section 5. Our hypotheses and the difference between the AER and ASPR criteria are illustrated with examples in section 6. We conclude in section 7 with some general remarks.

2. The optimal control problem. The control model that we are concerned with can be described by

$$\{S, A(i), q(j|i,a), r(i,a),\ i,j \in S\}, \tag{2.1}$$

where S is the *state space*; $A(i)$ is a set of *admissible actions* at state $i \in S$; $q(j|i,a)$ with $i,j \in S$ and $a \in A(i)$ are the system's *transition rates*; and $r(i,a)$ with $i \in S$ and $a \in A(i)$ are the *reward rates*. Let $K := \{(i,a) : i \in S,\ a \in A(i)\}$ be the set of all state-action pairs.

In this paper we assume that S is denumerable and in fact we write it as the set of nonnegative integers, i.e., $S = \{0, 1, 2, \ldots\}$. Furthermore, we assume that for each $i \in S$ the set $A(i)$ is a Borel space endowed with the Borel σ-algebra $\mathcal{B}(A(i))$. The transition rates $q(j|i,a)$ in (2.1) satisfy $q(j|i,a) \geq 0$ for all $(i,a) \in K$ and $j \neq i$. Moreover, we assume that the matrix $[q(j|i,a)]$ with (i,j)-element $q(j|i,a)$

is *conservative*, i.e.,

$$\sum_{j\in S} q(j|i,a) = 0 \quad \forall (i,a) \in K,$$

and *stable*, which means that

$$q(i) := \sup_{a\in A(i)} q_i(a) < \infty \quad \forall i \in S,$$

where $q_i(a) := -q(i|i,a) \geq 0$, for all $(i,a) \in K$. In addition, $q(j|i,a)$ is measurable in $a \in A(i)$ for each fixed $i, j \in S$.

Finally, the function $r(i,a)$ on K is a real-valued reward rate, and $r(i,a)$ is assumed to be measurable in $a \in A(i)$ for each fixed $i \in S$. (As $r(i,a)$ is allowed to take positive and negative values, it can be interpreted as a *cost rate* rather than a "reward" rate.)

We first introduce randomized Markov policies.

DEFINITION 2.1 (randomized Markov policies). *A randomized Markov policy is a function $\pi_t(B|i)$ that satisfies the following conditions:*

(1) *for each $i \in S$ and $B \in \mathcal{B}(A(i))$, the mapping $t \mapsto \pi_t(B|i)$ is Borel measurable on $[0,\infty)$, and*

(2) *for each $i \in S$ and $t \geq 0$, $B \mapsto \pi_t(B|i)$ is a probability measure on $\mathcal{B}(A(i))$.*

Let $A := \bigcup_{i\in S} A(i)$. A (deterministic) stationary policy is a function $f : S \to A$ such that $f(i)$ is in $A(i)$ for all $i \in S$.

Let Φ be the set of all randomized Markov policies and let F be the set of all stationary policies. Note that a function $f \in F$ can be viewed as a function $\pi_t(B|i) \in \Phi$ for which, for all $t \geq 0$ and $i \in S$, $\pi_t(\cdot|i)$ is the Dirac measure at $f(i)$. Thus, $F \subset \Phi$. We will write a randomized Markov policy $\pi_t(B|i)$ in Φ simply as (π_t). The subscript "t" in π_t indicates the possible dependence on time; it will be dropped for simplicity when there is no confusion.

For each $(\pi_t) \in \Phi$, the associated transition and reward rates are defined, respectively, as follows:

$$q(j|i,\pi_t) := \int_{A(i)} q(j|i,a)\pi_t(da|i) \quad \text{for } i,j \in S \text{ and } t \geq 0, \tag{2.2}$$

$$r(i,\pi_t) := \int_{A(i)} r(i,a)\pi_t(da|i) \quad \text{for } i \in S \text{ and } t \geq 0. \tag{2.3}$$

Obviously, the transition rate $q(j|i,\pi_t)$ and reward rate $r(i,\pi_t)$ can depend on time t if π is not stationary. When $\pi = f \in F$, we write $q(j|i,\pi_t)$ and $r(i,\pi_t)$ as $q(j|i,f(i))$ and $r(i,f(i))$, respectively.

For each $\pi := (\pi_t) \in \Phi$, let $Q(\pi_t) := [q(j|i,\pi_t)]$ with $t \geq 0$ be the transition rate matrices. Any (possibly substochastic and nonhomogeneous) transition function $\tilde{p}(s,i,t,j,\pi)$ such that

$$\lim_{\gamma\to 0^+} \frac{\tilde{p}(t,i,t+\gamma,j,\pi) - \delta_{ij}}{\gamma} = q(j|i,\pi_t) \quad \forall i,j \in S \text{ and } t \geq 0$$

is called a Q-*process* with the transition rate matrices $Q(\pi_t)$, where δ_{ij} is the Kronecker delta. To guarantee the existence of such a Q-process, we now define the class of admissible policies.

DEFINITION 2.2 (admissible policies). *A randomized Markov policy* (π_t) *in* Φ *is said to be admissible if* $q(j|i,\pi_t)$ *is continuous in* $t \geq 0$ *for each fixed* $i, j \in S$. *We denote by* Π *the class of all admissible policies.*

Π is *nonempty* because it contains F. Moreover, as shown in Example 6.3 below, Π contains a randomized Markov policy which is *not* in F.

On the other hand, $Q(\pi_t)$ is also conservative and stable, i.e.,

$$q_i(\pi_t) := -q(i|i,\pi_t) = \sum_{j\neq i} q(j|i,\pi_t) < \infty \quad \forall i \in S \text{ and } t \geq 0.$$

Hence, for each $\pi \in \Pi$, the existence of a Q-process such as the *minimum* Q-process denoted by $p^{\min}(s,i,t,j,\pi)$ (i.e., $p^{\min}(s,i,t,j,\pi) \leq \tilde{p}(s,i,t,j,\pi)$ for any Q-process $\tilde{p}(s,i,t,j,\pi)$) is guaranteed but is not necessarily regular; that is, we might have $\sum_{j\in S} p^{\min}(s,i,t,j,\pi) < 1$ for some $i \in S$ and $t \geq s \geq 0$ (see [13] or Theorem 4.2.6 in [2]).

To ensure the regularity of a Q-process, we use the following ergodicity conditions.

Assumption A. There exist a sequence $\{S_n,\ n \geq 1\}$ of subsets of S, a nondecreasing function $w \geq 1$ on S, and two constants $c > 0$ and $b \geq 0$, such that

(1) $\sup_{i\in S_n} q(i) < \infty$ for each $n \geq 1$, and $S_n \uparrow S$ in the sense of convergence of a set sequence;

(2) $\lim_{n\to\infty}[\inf_{j\notin S_n} w(j)] = +\infty$;

(3) $\sum_{j\in S} q(j|i,a)w(j) \leq -cw(i) + b\delta_{0i} \forall\ (i,a) \in K$; and

(4) for each $f \in F$, the minimum Q-process $p^{\min}(s,i,t,j,f)$ is *monotone*, i.e.,

$$\sum_{j\geq k} q(j|i,f(i)) \leq \sum_{j\geq k} q(j|i+1,f(i+1)) \quad \forall i,k \in S \text{ with } k \neq i+1,$$

and *irreducible*, i.e., for each pair of states i and j, either $q(j|i,f(i)) > 0$, or there are an integer l (which may depend on i, j, and f) and l states $i_1, i_2, \ldots, i_l$ with $i \neq i_1$, $j \neq i_l$, $i_{k-1} \neq i_k$, $k = 2, \ldots, l$, such that

$$q(i_1|i,f(i))q(i_2|i_1,f(i_1)) \cdots q(i_l|i_{l-1},f(i_{l-1}))q(j|i_l,f(i_l)) > 0.$$

LEMMA 2.3. (a) *If Assumptions* A(1), A(2), *and* A(3) *hold, then for each* $\pi = (\pi_t) \in \Pi$ *the corresponding Q-process with transition rate matrices* $Q(\pi_t)$ *is regular; that is,*

$$\sum_{j\in S} p^{\min}(s,i,t,j,\pi) = 1 \quad \forall i \in S \text{ and } t \geq s \geq 0.$$

(b) *If Assumption* A *holds, then for each* $f \in F$ *the corresponding Q-process with transition rate matrices* $[Q(j|i,f(i))]$ *is ergodic, and its unique invariant probability measure* μ_f *(with* $\mu_f(i) > 0$ *for all* $i \in S$*) can be determined by the equation*

$$\sum_{i\in S} \mu_f(i)q(j|i,f(i)) = 0 \quad \forall j \in S. \tag{2.4}$$

Moreover, for each $i \in S$ *and* $t \geq 0$

$$\left|\sum_{j\in S} p^{\min}(0,i,t,j,f)h(j) - \mu_f(h)\right| \leq 2e^{-ct}\left[w(i) + \frac{b}{c}\right] \leq 2e^{-ct}\left(1 + \frac{b}{c}\right)w(i) \tag{2.5}$$

for any function h on S such that $|h| \le w$, where $\mu_f(h) := \sum_{j\in S} h(j)\mu_f(j)$.

Proof. (a) Under Assumptions A(1)–A(3), by Theorem 3.1 in [17] we see that (a) is true.

(b) By (a) and Proposition 5.4.1 in [2], we see that (2.4) is true. Moreover, from the proof of (3.9) in [30] we see that the condition (2.1) in [30] is not required for Theorem 2.2(ii) in [30]. Thus, by (3.9) in [30] and Assumption A we see that (2.5) is also true. □

Under Assumptions A(1)–A(3), Lemma 2.3 shows that for each $\pi = (\pi_t) \in \Pi$ a Q-process with transition rate matrices $Q(\pi_t)$ is regular. Thus, under Assumption A, we will denote by $\{x(t,\pi)\}$ the associated right-continuous Markov chain with values in S, and write the regular Q-process $p^{\min}(s,i,t,j,\pi)$ simply as $p(s,i,t,j,\pi)$. Furthermore, for each initial state $i \in S$ at time $s = 0$, we denote by $(\Omega, \mathcal{F}, P_i^\pi)$ the probability measure space determined by $p(s,i,t,j,\pi)$, by E_i^π the corresponding expectation operator, and by $x(t,\pi)(e)$ the value of $x(t,\pi)$ at $e \in \mathcal{F}$.

Remark 2.4. (a) For the case where $\sup_{i\in S} q(i) < \infty$ (see, for instance, [7, 26, 31, 35, 39]), Assumptions A(1) and A(2) are not required because they are used only to guarantee the regularity of a Q-process. For the case of *unbounded* transition rates (e.g., [18, 19]), the conditions for a Q-process to be regular are usually imposed on both the possibly *nonhomogeneous* minimum Q-processes and the transition rates. Hence, our Assumptions A(1)–A(3) are quite different from those in [18, 19].

(b) Assumptions A(1)–A(3) are an extension of both the "drift condition" in [30] and the hypotheses of Corollary 2.2.16 in [2] for a *homogeneous* Q-process to be regular. Assumption A(4) is a variant of the monotonicity conditions in Theorem 7.3.4 and the irreducibility conditions in Proposition 5.3.1 in [2].

(c) It should be mentioned that if there is a set $\bar{S}$ of transient states which is independent of stationary policies, then $\mu_f(i) = 0$ for each $i \in \bar{S}$ and $f \in F$. In this case, Lemma 3.4 below may *not* hold because its proof uses the result $\mu_f(i) > 0$ for all $i \in S$.

Now we define the ASPR criterion $V_{sp}(\cdot,\cdot)$ as follows: for each $\pi = (\pi_t) \in \Pi$ and $i \in S$

$$V_{sp}(\pi,i) := \limsup_{T\to\infty} \frac{1}{T}\left[\int_0^T r(x(t,\pi),\pi_t)dt\right], \tag{2.6}$$

where the subscript "sp" stands for "sample-path." Note that $V_{sp}(\pi,i)$ has been defined by the so-called *sample-path rewards* $r(x(t,\pi),\pi_t)$; therefore, it is a *random variable* rather than a number as in the AER-criterion defined as

$$\bar{V}(\pi,i) := \limsup_{T\to\infty} \frac{1}{T}\left[\int_0^T E_i^\pi r(x(t,\pi),\pi_t)dt\right] \tag{2.7}$$

(see [4, 7, 16, 19, 20, 21, 26, 27, 31, 35, 39], for instance). Thus, the following definition of optimal policies for the ASPR criterion is different from that for the AER criterion.

Definition 2.5. *For a given $\epsilon \ge 0$, a policy $\pi^* \in \Pi$ is said to be ϵ-ASPR-optimal if there exists a constant g^* such that*

$$P_i^{\pi^*}(V_{sp}(\pi^*,i) \ge g^* - \epsilon) = 1 \quad \text{and} \quad P_i^{\pi}(V_{sp}(\pi,i) \le g^*) = 1 \quad \forall i \in S \text{ and } \pi \in \Pi.$$

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

The main goal of this paper is to give conditions on the primitive data in (2.1) that ensure the existence of an ASPR-optimal stationary policy.

3. Preliminaries. In this section we present some preliminary facts that are needed to prove our main results.

Let $w \geq 1$ be the function in Assumption A. Following the concept of a weighted supremum norm introduced by Lippman [28] and widely used by many authors (e.g., [23, p. 2]), we define the weighted supremum norm $\|v\|_w$ for a real-valued functions v on S by

$$\|v\|_w := \sup_{i \in S}[w(i)^{-1}|v(i)|]$$

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

LEMMA 3.1. *Let $\bar{w}$ be any nonnegative function on S, and $\bar{c}$, $\bar{b}$ two constants such that $\bar{b} \geq 0$ and $\bar{c} \neq 0$. Then, for each $\pi = (\pi_t) \in \Pi$, the following statements are equivalent:*

(a) $\sum_{j \in S} p^{\min}(s, i, t, j, \pi)\bar{w}(j) \leq e^{-\bar{c}(t-s)}\bar{w}(i) + \frac{\bar{b}}{\bar{c}}[1 - e^{-\bar{c}(t-s)}]$ *for all $i \in S$ and $t \geq s \geq 0$;*

(b) $\sum_{j \in S} q(j|i, \pi_t)\bar{w}(j) \leq -\bar{c}\bar{w}(i) + \bar{b}$ *for all $i \in S$ and $t \geq 0$.*

Proof. See Lemma 3.2 in [16]. □

It should be noted that in Lemma 3.1, Assumption A is *not* required.

To establish the so-called optimality equation, we will use a *policy iteration algorithm* instead of the *vanishing discount approach* in [16, 19, 21, 26]. To state the policy iteration algorithm, in addition to Assumption A we also need the following standard continuity-compactness conditions (Assumption B); see, for instance, [3, 19, 23, 31, 35] and their references.

Assumption B. (1) For each $i \in S$, $A(i)$ is compact.

(2) $r(i, a)$ and $q(j|i, a)$ are continuous in $a \in A(i)$ for each fixed $i, j \in S$.

(3) The function $\sum_{j \in S} q(j|i, a)w(j)$ is continuous in $a \in A(i)$ for each fixed $i \in S$.

(4) There exists a positive constant M such that $|r(i, a)| \leq Mw(i)$ for all $i \in S$ and $a \in A(i)$.

In the spirit of the potential concept in [8, 7], for a given $f \in F$ and the corresponding unique invariant probability measure μ_f, we define the *potential*

$$u(f, i) := \int_0^\infty [E_i^f r(x(t, f), f(x(t, f))) - g(f)]dt \quad \forall\, i \in S, \tag{3.1}$$

where the constant $g(f)$ is defined as

$$g(f) := \sum_{j \in S} r(j, f(j))\mu_f(j). \tag{3.2}$$

LEMMA 3.2. *Let Assumptions* A *and* B(4) *hold. Then*

(a) *$g(f)$ and $\|u(f, \cdot)\|_w$ are both bounded in $f \in F$,*

(b) *the Poisson equation $g(f) = r(i, f(i)) + \sum_{j \in S} q(j|i, f(i))u(f, j)$ holds for all $i \in S$ and $f \in F$.*

Proof. By (3.1) and (2.5) we see that $\|u(f, \cdot)\|_w$ is bounded in $f \in F$. With the constants M, c, and b as in Assumptions A and B(4), by Lemma 3.1, (3.1), (2.5), and (2.7), we have $|\bar{V}(f, \cdot)| = |g(f)| \leq \frac{Mb}{c}$ for all $f \in F$, and so (a) follows. Obviously, (b) follows from Lemma 5.1 in [16]. □

Under Assumptions A and B, we now state the policy iteration algorithm.

POLICY ITERATION ALGORITHM 3.1.

Step I. Take $n = 0$ and $f_n \in F$.

Step II. Solve (2.4) for μ_{f_n} and then calculate $u(f_n, \cdot)$ and $g(f_n)$ as in (3.1) and (3.2).

Step III. Define a new stationary policy f_{n+1} in the following way:

Set $f_{n+1}(i) := f_n(i)$ for all $i \in S$ for which

$$r(i, f_n(i)) + \sum_{j \in S} q(j|i, f_n(i))u(f_n, j) = \max_{a \in A(i)} \left\{ r(i, a) + \sum_{j \in S} q(j|i, a)u(f_n, j) \right\}; \tag{3.3}$$

otherwise (i.e., when (3.3) does not hold), choose $f_{n+1}(i) \in A(i)$ such that

$$\begin{aligned} &r(i, f_{n+1}(i)) + \sum_{j \in S} q(j|i, f_{n+1}(i))u(f_n, j) \\ &\quad = \max_{a \in A(i)} \left\{ r(i, a) + \sum_{j \in S} q(j|i, a)u(f_n, j) \right\}. \end{aligned} \tag{3.4}$$

Step IV. If $f_{n+1}(i)$ satisfies (3.3) for all $i \in S$, then stop because (by Theorem 4.1 below) f_{n+1} is ASPR-optimal; otherwise, replace f_n with f_{n+1} and go back to *Step* II.

Finally, to prove the existence of an ASPR-optimal stationary policy, in addition to Assumptions A and B we propose the following conditions.

Assumption C. There exist nonnegative functions $w_k^* \geq 1$ on S as well as constants $c_k^* > 0$, $b_k^* \geq 0$, and $M_k^* > 0$ $(k = 1, 2)$ such that, for each $i \in S$ and $a \in A(i)$,

(1) $w^2(i) \leq M_1^* w_1^*(i)$ and $\sum_{j \in S} q(j|i, a)w_1^*(j) \leq -c_1^* w_1^*(i) + b_1^*$, and

(2) $[q(i)w(i)]^2 \leq M_2^* w_2^*(i)$ and $\sum_{j \in S} q(j|i, a)w_2^*(j) \leq -c_2^* w_k^*(i) + b_2^*$.

Remark 3.3. (a) Assumption C allows us to use the martingale stability theorem; see Lemma 3.11 in [22], for instance. However, it is not required when a solution u^* in (4.1) below and the transition rates are both uniformly bounded.

(b) Assumption C(2) is slightly different from Assumption B(4) in [16], but all conclusions in [16] still hold after Assumption B(4) in [16] is replaced by Assumption C(2) here.

For each $n \geq 1$, take f_n as the policy obtained in the policy iteration algorithm 3.1, and for each $i \in S$ let

$$\varepsilon(f_n, i) := r(i, f_n(i)) + \sum_{j \in S} q(j|i, f_n(i))u(f_{n-1}, j) - g(f_{n-1}). \tag{3.5}$$

LEMMA 3.4. *Let Assumptions* A, B, *and* C(2) *hold. Then* $g(f_{n+1}) > g(f_n)$ *when* $f_{n+1} \neq f_n$, *and for each* $i \in S$, $\varepsilon(f_n, i) \to 0$ *as* $n \to \infty$.

Proof. As in the proof of Theorem 5.2 and Lemma 5.3 in [16], by Lemma 3.2 above we obtain Lemma 3.4. □

Lemma 3.4 will be used to establish the optimality equation (4.1) below.

4. The existence of ASPR-optimal stationary policies. In this section, we state and prove our main result, Theorem 4.1.

THEOREM 4.1. *Under Assumptions* A, B, *and* C, *the following statements hold.*

(a) *There exist a unique constant* g^*, *a function* $u^* \in B_w(S)$, *and a stationary policy* $f^* \in F$ *satisfying the optimality equation*

$$g^* = r(i, f^*(i)) + \sum_{j \in S} q(j|i, f^*(i))u^*(j)$$

$$= \max_{a\in A(i)}\left\{r(i,a)+\sum_{j\in S}q(j|i,a)u^*(j)\right\} \quad \forall i\in S. \tag{4.1}$$

(b) *The policy f^* in* (a) *is ASPR-optimal, and $P_i^{f^*}(V_{sp}(f^*,i)=g^*)=1$ for all $i\in S$.*

(c) *A policy f in F is ASPR-optimal if and only if it realizes the maximum of* (4.1).

(d) *For given $\epsilon\geq 0$ and $f\in F$, if there is a function $\bar{u}\in B_w(S)$ such that*

$$g^*\leq r(i,f(i))+\sum_{j\in S}q(j|i,f(i))\bar{u}(j)+\epsilon \quad \forall i\in S,$$

then f is ϵ-ASPR-optimal.

Proof. (a) Let $\{f_n\}$ be the sequence of the stationary policies obtained by the policy iteration algorithm 3.1. By Assumption B(1) and the Tichonoff theorem, the policy class F is compact. Thus, by Lemma 3.2(a), there exist a subsequence $\{f_{n_k}\}$ of $\{f_n\}$ and $u^*\in B_w(S)$ such that for each $i\in S$

$$\lim_{k\to\infty}u(f_{n_k},i)=u^*(i), \quad \lim_{k\to\infty}f_{n_k}(i)=:f^*(i), \quad \text{and} \quad \lim_{k\to\infty}g(f_{n_k})=:g^*. \tag{4.2}$$

On the other hand, by Lemmas 3.2(b), (3.4), and (3.5), we have

$$\begin{aligned} g(f_{n_k}) &= r(i,f_{n_k}(i))+\sum_{j\in S}q(j|i,f_{n_k}(i))u(f_{n_k},j)\\ &= \max_{a\in A(i)}\left\{r(i,a)+\sum_{j\in S}q(j|i,a)u(f_{n_k},j)\right\}-\varepsilon(f_{n_k+1},i)\\ &\geq r(i,a)+\sum_{j\in S}q(j|i,a)u(f_{n_k},j)-\varepsilon(f_{n_k+1},i) \quad \forall i\in S \text{ and } a\in A(i). \end{aligned} \tag{4.3}$$

Letting $k\to\infty$ in (4.3), by the "extension of Fatou's Lemma" 8.3.7 in [23] and our Lemma 3.4 and (4.2), we have

$$\begin{aligned} g^* &= r(i,f^*(i))+\sum_{j\in S}q(j|i,f^*(i))u^*(j)\\ &\geq r(i,a)+\sum_{j\in S}q(j|i,a)u^*(j) \quad \forall i\in S \text{ and } a\in A(i), \end{aligned}$$

and so

$$\begin{aligned} g^* &= r(i,f^*(i))+\sum_{j\in S}q(j|i,f^*(i))u^*(j)\\ &= \max_{a\in A(i)}\left\{r(i,a)+\sum_{j\in S}q(j|i,a)u^*(j)\right\} \quad \forall i\in S, \end{aligned}$$

which gives (4.1). Moreover, the proof of the uniqueness of the constant g^* satisfying (4.1) follows from Theorem 4.1(b) in [16] and Remark 3.3(b).

(b) To prove (b), for each $i\in S$, $\pi=(\pi_t)\in\Pi$, and $t\geq 0$, let

$$\Delta(i,\pi_t):=r(i,\pi_t)+\sum_{j\in S}q(j|i,\pi_t)u^*(j)-g^*, \tag{4.4}$$

$$\mathcal{F}_t(\pi) := \sigma\{x(s,\pi),\ 0 \le s \le t\},$$
$$g(i,\pi_t) := \sum_{j\in S} q(j|i,\pi_t)u^*(j). \tag{4.5}$$

In particular, let $\Delta(i,f(i)) =: r(i,f(i)) + \sum_{j\in S} q(j|i,f(i))u^*(j) - g^*$ for all $f \in F$.

We now define a (continuous-time) stochastic process,

$$M(t,\pi) := \int_0^t g(x(y,\pi),\pi_y)dy - u^*(x(t,\pi)) \quad \text{for } t \ge 0. \tag{4.6}$$

Then $\{M(t,\pi), \mathcal{F}_t(\pi),\ t \ge 0\}$ is a P_i^π-martingale in continuous-time; that is,

$$E_i^\pi[M(t,\pi)|\mathcal{F}_s(\pi)] = M(s,\pi) \quad \forall t \ge s \ge 0. \tag{4.7}$$

Indeed, for each $t \ge s \ge 0$, by (4.6) and the Markov property we have

$$\begin{aligned} E_i^\pi[M(t,\pi)|\mathcal{F}_s(\pi)] &= M(s,\pi) + E_i^\pi\left[\int_s^t g(x(y,\pi),\pi_y)dy|\mathcal{F}_s(\pi)\right] \\ &\quad + u^*(x(s,\pi)) - E^\pi_{x(s,\pi)}u^*(x(t,\pi)). \end{aligned} \tag{4.8}$$

Since $u^* \in B_w(S)$ (by (4.2) and Lemma 3.2(a)), it follows from Assumption A(3) that

$$\begin{aligned} \left|\sum_{j\in S} q(j|i,a)u^*(j)\right| &\le \|u^*\|_w\left[\sum_{j\in S} q(j|i,a)w(j) - 2q(i|i,a)w(i)\right] \\ &\le \|u^*\|_w[-cw(i) + b + 2q(i)w(i)] \\ &\le \|u^*\|_w[b + 2q(i)w(i)] \end{aligned} \tag{4.9}$$

for all $a \in A(i)$ and $i \in S$. Therefore, by (4.5) and (2.2) we obtain

$$|g(i,\pi_y)| \le \|u^*\|_w[b + 2q(i)w(i)] \quad \forall y \ge 0 \text{ and } i \in S. \tag{4.10}$$

On the other hand, by the Markov property we have

$$E_i^\pi\left[\int_s^t g(x(y,\pi),\pi_y)dy|\mathcal{F}_s(\pi)\right] = E^\pi_{x(s,\pi)}\left[\int_s^t g(x(y,\pi),\pi_y)dy\right],$$

which together with (4.10), Assumption C(2), Lemma 3.1, and Fubini's theorem gives

$$E_i^\pi\left[\int_s^t g(x(y,\pi),\pi_y)dy|\mathcal{F}_s(\pi)\right] = \int_s^t\left[E^\pi_{x(s,\pi)}g(x(y,\pi),\pi_y)\right]dy. \tag{4.11}$$

From Lemma 2.1(b) in [21] and (4.10) in [16] about the *extended generator* of a possibly nonhomogeneous continuous-time Markov process, by (4.11) and (4.5) we obtain

$$E_i^\pi\left[\int_s^t g(x(y,\pi),\pi_y)dy|\mathcal{F}_s(\pi)\right] = E^\pi_{x(s,\pi)}u^*(x(t,\pi)) - u^*(x(s,\pi)),$$

which together with (4.8) gives (4.7).

It follows from (4.7) that $\{M(n,\pi), \mathcal{F}_n(\pi),\ n \ge 1\}$ is also a P_i^π-martingale in discrete-time. Moreover, By Assumption C and Lemma 3.1 we have

$$E_i^\pi w_k^*(x(t,\pi)) \le w_k^*(i) + \frac{b_k^*}{c_k^*} \quad \forall t \ge 0 \text{ and } k = 1,2, \tag{4.12}$$

which together with (4.6), (4.10), the Hölder inequality, and Assumption C gives

$$
\begin{aligned}
&E_i^\pi[M(n+1,\pi)-M(n,\pi)]^2\\
&\quad= E_i^\pi\left[\int_n^{n+1} g(x(y,\pi),\pi_y)dy + u^*(x(n,\pi)) - u^*(x(n+1,\pi))\right]^2\\
&\quad\le 2E_i^\pi\left[\int_n^{n+1} g(x(y,\pi),\pi_y)dy\right]^2 + 2E_i^\pi[u^*(x(n+1,\pi)) - u^*(x(n,\pi))]^2\\
&\quad\le 2E_i^\pi\left[\int_n^{n+1} g^2(x(y,\pi),\pi_y)dy\right] \qquad \text{(by the Hölder inequality)}\\
&\qquad + 4\|u^*\|_w^2 E_i^\pi[w^2(x(n+1,\pi)) + w^2(x(n,\pi))]\\
&\quad\le 2E_i^\pi\left[\int_n^{n+1} \|u^*\|_w^2[b + 2q(x(y,\pi))w(x(y,\pi))]^2 dy\right] \quad \text{(by (4.10))}\\
&\qquad + 4M_1^*\|u^*\|_w^2 E_i^\pi[w_1^*(x(n+1,\pi)) + w_1^*(x(n,\pi))] \qquad \text{(by Assumption C(1))}\\
&\quad\le 4\|u^*\|_w^2 E_i^\pi\left[\int_n^{n+1} (b^2 + 4[q(x(y,\pi))w(x(y,\pi))]^2)dy\right]\\
&\qquad + 4M_1^*\|u^*\|_w^2 E_i^\pi[w_1^*(x(n+1,\pi)) + w_1^*(x(n,\pi))]\\
&\quad\le 4\|u^*\|_w^2 E_i^\pi\left[\int_n^{n+1} (b^2 + 4M_2^* w_2^*(x(y,\pi)))dy\right] \qquad \text{(by Assumption C(2))}\\
&\qquad + 4M_1^*\|u^*\|_w^2 E_i^\pi[w_1^*(x(n+1,\pi)) + w_1^*(x(n,\pi))],
\end{aligned}
$$

which gives

$$
\begin{aligned}
&E_i^\pi[M(n+1,\pi)-M(n,\pi)]^2\\
&\qquad\le 16\|u^*\|_w^2\left[b^2 + M_2^*\left(w_2^*(i) + \frac{b_2^*}{c_2^*}\right) + M_1^*\left(w_1^*(i) + \frac{b_1^*}{c_1^*}\right)\right] \quad \text{(by (4.12))}.
\end{aligned} \tag{4.13}
$$

This means that $E_i^\pi[M(n+1,\pi)-M(n,\pi)]^2$ is bounded in $n \ge 1$. Thus, by the martingale stability theorem (e.g., [22, p. 105], or Remark 11.2.6 in [23], for instance), we have

$$
\lim_{n\to\infty}\frac{M(n,\pi)}{n} = 0 \quad \text{a.s.} - P_i^\pi. \tag{4.14}
$$

On the other hand, for any $T \ge 1$, let $[T]$ be the unique integer such that $[T] \le T < [T]+1$. By (4.6) we have

$$
\frac{M(T,\pi)}{T} = \frac{[T]}{T}\left(\frac{M([T],\pi)}{[T]} + \frac{\int_{[T]}^T g(x(y,\pi),\pi_y)dy}{[T]} - \frac{u^*(x(T,\pi))}{[T]} + \frac{u^*(x([T],\pi))}{[T]}\right). \tag{4.15}
$$

Moreover, for any arbitrary $\epsilon > 0$, as in the proof of (4.13), by the Chebyshev's inequality we have

$$
\begin{aligned}
P_i^\pi\left(\frac{\left|\int_{[T]}^T g(x(y,\pi),\pi_y)dy\right|}{[T]} > \epsilon\right) &\le \frac{E_i^\pi\left[\int_{[T]}^T |g(x(y,\pi),\pi_y)|dy\right]^2}{\epsilon^2[T]^2}\\
&\le \frac{16\|u^*\|_w^2\left[b^2 + M_2^*\left(w_2^*(i) + \frac{b_2^*}{c_2^*}\right)\right]}{\epsilon^2[T]^2}.
\end{aligned} \tag{4.16}
$$

Since $\sum_{[T]=1}^{\infty} \frac{1}{[T]^2} < \infty$, by (4.16) and the Borel–Cantelli lemma, we have

$$P_i^\pi \left(\limsup_{[T]} \left\{ \frac{\left| \int_{[T]}^T g(x(y,\pi), \pi_y) dy \right|}{[T]} > \epsilon \right\} \right) = 0.$$

Now let

$$E_{[T]} := \left\{ \frac{|\int_{[T]}^T g(x(y,\pi), \pi_y) dy|}{[T]} > \epsilon \right\} \in \mathcal{F},$$

$E := \limsup_{[T]} E_{[T]} \in \mathcal{F}$, and $E^c := \Omega - E$ being the complement of set E. Then $P_i^\pi(E^c) = 1$. Let $e \in E^c$, which means that e is in finitely many sets $E_{[T]}$. So there exists an integer $N_0(e)$ (depending on e) such that $e \notin E_{[T]}$ for all $[T] \geq N_0(e)$, i.e.,

$$\frac{\left| \int_{[T]}^T g(x(y,\pi)(e), \pi_y) dy \right|}{[T]} \leq \epsilon \quad \forall\ [T] \geq N_0(e) \text{ and } e \in E^c,$$

which together with $P_i^\pi(E^c) = 1$ yields

$$\lim_{[T]\to\infty} \frac{\int_{[T]}^T g(x(y,\pi), \pi_y) dy}{[T]} = 0 \quad \text{a.s.} - P_i^\pi. \tag{4.17}$$

Similarly, we have

$$\lim_{[T]\to\infty} \frac{u^*(x(T,\pi))}{[T]} = \lim_{[T]\to\infty} \frac{u^*(x([T],\pi))}{[T]} = 0 \quad \text{a.s.} - P_i^\pi. \tag{4.18}$$

Since $\lim_{T\to\infty} \frac{[T]}{T} = 1$, by (4.14), (4.15), (4.17), and (4.18), we have

$$\lim_{T\to\infty} \frac{M(T,\pi)}{T} = 0 \quad \text{a.s.} - P_i^\pi. \tag{4.19}$$

By (4.4)–(4.6) it follows that

$$M(t,\pi) = -\int_0^t r(x(y,\pi), \pi_y) dy + \int_0^t \Delta(x(y,\pi), \pi_y) dy - u^*(x(t,\pi)) + tg^*. \tag{4.20}$$

By (4.1), (4.4), (2.2), and (2.3), we have $\Delta(i, \pi_t) \leq 0$ and $\Delta(i, f^*(i)) = 0$ for all $t \geq 0$ and $i \in S$. Thus, by (4.18), (4.19), and (4.20) we obtain

$$P_i^\pi(V_{sp}(\pi, i) \leq g^*) = 1 \ \text{ and} \tag{4.21}$$

$$P_i^{f^*}(V_{sp}(f^*, i) = g^*) = 1, \tag{4.22}$$

which, together with the arbitrariness of π and i, give (b).

(c) By (b), it suffices to prove that $f(\in F)$ realizes the maximum of (4.1) if f is SPAR-optimal. Now suppose that f is SPAR-optimal but does not realize the maximum of (4.1). Then there exist some $i_0 \in S$ and a constant $\alpha(i_0, f) > 0$ (depending on i_0 and f) such that

$$g^* \geq [r(i, f(i)) + \alpha(i_0, f)\delta_{i_0 i}] + \sum_{j\in S} q(j|i, f(i)) u^*(j) \quad \forall i \in S. \tag{4.23}$$

On the other hand, since f is SPAR-optimal, by (b) and (4.21) we have $V_{sp}(f,i) = g^*$ a.s. for all $i \in S$. Moreover, as in the proof of (4.22), from Lemma 3.2(b) we also have $V_{sp}(f,i) = g(f)$ a.s., and so

$$g^* = g(f) = \sum_{j\in S} \mu_f(j) r(j, f(j)). \tag{4.24}$$

Also, as in the proof of (4.12) in [16], by (4.23) and (4.24) as well as (2.7) we obtain

$$g^* \geq \sum_{j\in S} \mu_f(j)[r(j,f(j)) + \alpha(i_0,f)\delta_{i_0 j}] = g^* + \mu_f(i_0)\alpha(i_0,f),$$

which gives a contradiction because $\mu_f(i_0)$ and $\alpha(i_0,f)$ are both positive.

(d) Let $\Delta_{\bar{u}}(i,f(i)) := r(i,f(i)) + \sum_{j\in S} q(j|i,f(i))\bar{u}(j) - g^*$. Then, $\Delta_{\bar{u}}(i,f(i)) \geq -\epsilon$ for all $i \in S$. Thus, as in the proof of (4.21), we have

$$P_i^f(V_{sp}(f,i) \geq g^* - \epsilon) = 1,$$

which together with (b) gives (d). □

Theorem 4.1 is an important result: part (a) establishes the optimality equation (4.1) and the existence of a so-called *canonical* policy f^*, whereas part (b) further shows that the canonical policy f^* is ASPR-optimal.

Remark 4.2. (a) Under Assumptions A, B, and C(2) only, from the proof of Theorem 4.1 here and Theorem 4.1 in [16] we see that the canonical policy f^* in Theorem 4.1(a) is also optimal for the AER criterion. However, it is shown that an optimal stationary policy for the AER criterion may *not* be canonical [18]. Therefore, it is natural to *guess* that an ASPR-optimal stationary policy may *not* be canonical either. An attempt to answer this problem faces significant technical difficulties, and the problem remains unsolved to this date.

(b) From the proof of Theorem 4.1(b) and (c) we see that both Assumptions C(1) and C(2) are indeed required for the ASPR criterion. That is because (i) the proof of Theorem 4.1(b) and (c) uses the estimates in (4.13) and (4.16), and (ii) the proof of (4.13) and (4.16) is based on both Assumptions C(1) and C(2); see the proof of (4.12) and (4.24) (In the proof of the "if" part of Theorem 4.1(c), we cannot obtain (4.24) by the dominated convergence theorem because $V_{sp}(i,f)$ is defined via "limsup" instead of "lim.")

(c) To establish the optimality equation (4.1), we have used the *policy iteration algorithm* 3.1, instead of the "vanishing discount factor method" used in [16, 18, 19, 21, 26], for instance. It should be noted that our approach is *direct* because it does not require any result about discounted continuous-time MDPs. This is by way of the same logic introduced in [25] for *discrete-time* unichain MDPs. (A similar approach is adopted for discrete-time general MDPs with finite state and action sets in [9, 38] by using simple algebra and properties of the Cesaro-limit of a transition probability matrix and in [12, 31] by using vanishing discount factors.)

(d) We can also prove Theorem 4.1 by using the "vanishing discount factor method." More precisely, under Assumptions A and B, we can (i) establish the average optimality inequalities by using the α-discounted optimality equation in [17], (ii) obtain the optimality equation, and (iii) prove the existence of ASPR-optimal stationary policies under the additional Assumption C. However, this vanishing factor method needs *additional* results about discounted continuous-time MDPs in [17].

When the transition and reward rates are both *uniformly bounded*, we need to impose conditions only on the *embedded* Markov chains to guarantee the existence of SPAR-optimal stationary policies. This is stated in the following corollary.

COROLLARY 4.3. *Suppose the following conditions* (1)–(3) *are satisfied.*

(1) $\|q\| := \sup_{i\in S} q(i) < \infty$, $\|r\| := \sup_{i\in S, a\in A(i)} |r(i,a)| < \infty$.

(2) *For each* $i \in S$, $A(i)$ *is compact; and* $r(i,a)$ *and* $q(j|i,a)$ *are continuous in* $a \in A(i)$ *for each fixed* $i, j \in S$.

(3) *Either* $\inf_{i\neq j_0, a\in A(i)} q(j_0|i,a) > 0$ *for some* $j_0 \in S$; *or* $\sum_{j\in S} \sup_{i\in S, a\in A(i)} (\frac{q(j|i,a)}{\|q\|} + \delta_{ij}) < 2$.

Then, the following results hold.

(a) *There exists an ASPR-optimal stationary policy.*

(b) *For each* $\epsilon > 0$, *an* ϵ*-ASPR-optimal stationary policy can be obtained in a finite number of steps of the policy iteration algorithm* 3.1.

Proof. Define maps T_k on the set $M(S)$ of *bounded functions* on S as

$$T_k u(i) := \sup_{a\in A(i)} \left\{ \frac{r(i,a)}{\|q\|+1} + \sum_{j\in S} \left[\left(\frac{q(j|i,a)}{\|q\|+1} + \delta_{ij} \right) - \mu_k(j) \right] u(j) \right\} \tag{4.25}$$

for all $i \in S$, $u \in M(S)$, and $k = 1, 2$, where the measures μ_k on S are given by

$$\mu_1(j) := \inf_{i\in S, a\in A(i)} \left[\frac{q(j|i,a)}{\|q\|+1} + \delta_{ij} \right] \text{ and}$$

$$\mu_2(j) := \sup_{i\in S, a\in A(i)} \left[\frac{q(j|i,a)}{\|q\|+1} + \delta_{ij} \right] \text{ for } j \in S,$$

which correspond to the first and second hypotheses in the condition (3), respectively. Thus, the maps T_1 and T_2 are both contractive with contraction factors β_1 and β_2, respectively, where

$$\beta_1 := 1 - \mu_1(S) \in (0,1) \text{ and } \beta_2 := \mu_2(S) - 1 \in (0,1). \tag{4.26}$$

Hence, the Banach's fixed point theorem gives the existence of $u^* \in M(S)$, $f^* \in F$ and a unique constant g^* satisfying (4.1). Then, as in the proof of Theorem 4.1(a) and (d), we see that Corollary 4.3 is true. □

Remark 4.4. (a) The two sets in the condition (3) in Corollary 4.3 are variants of the ergodicity condition in [22] for discrete-time MDPs, and each set implies that the embedded chain with the transition probability $(\frac{q(j|i,f(i))}{1+\|q\|} + \delta_{ij})$ has a unique invariant probability measure; see p. 56 in [22], for instance. The difference between the "monotonicity" condition in Assumption A(4) and the condition (3) in Corollary 4.3 can be shown by examples.

(b) Corollary 4.3 can also be obtained by using the uniformization method in [29, 31, 36] and the equivalence between continuous- and discrete-time MDPs in [31, 36, 37], as well as the results for discrete-time MDPs in [3, 10, 14, 15, 22, 24, 32, 34], for instance.

5. Algorithms. Following the procedure in the proof of Theorem 4.1, we now provide a policy iteration algorithm to obtain ASPR-optimal stationary policies.

PROPOSITION 5.1. *Suppose that Assumptions* A, B, *and* C *hold. Then any limit point* f^* *of the sequence* $\{f_n\}$ *obtained by the policy iteration Algorithm* 3.1 *is ASPR-optimal.*

Proof. The proposition follows directly from the proof of Theorem 4.1. □

Under the conditions in Corollary 4.3, we provide a *value iteration algorithm* to compute $\epsilon(>0)$-ASPR-optimal stationary policies. It should be mentioned that, as in the proof of Corollary 4.3, the choice of $k = 1$ (or 2) corresponds to the first (or second) hypothesis in the condition (3) in Corollary 4.3. Thus, we will understand that k in this algorithm is *fixed*.

VALUE ITERATION ALGORITHM 5.1.

Step I. For a fixed $\epsilon > 0$, take arbitrarily $u_0 \in M(S)$.

Step II. If $T_k u_0 = u_0$, then obtain a policy f (in F) satisfying

$$r(i, f(i)) + \sum_{j\in S} q(j|i, f(i))u_0(j) = \sup_{a\in A(i)} \left\{ r(i,a) + \sum_{j\in S} q(j|i,a)u_0(j) \right\} \quad \forall i \in S,$$

and f is ASPR-optimal (by Theorem 4.1), stop; otherwise, calculate a positive integer $N \geq \frac{1}{\beta_k} \ln \frac{\epsilon(1-\beta_k)}{4(1+\|q\|)\|u_1-u_0\|} + 1$ with β_k as in (4.26), and $u_N := T_k^N u_0 = T_k(T_k^{N-1}u_0)$ (by (4.25)).

Step III. Choose $f_\epsilon(i) \in A(i)$ such that for each $i \in S$

$$r(i, f_\epsilon(i)) + \sum_{j\in S} q(j|i, f_\epsilon(i))u_N(j) \geq \sup_{a\in A(i)} \left\{ r(i,a) + \sum_{j\in S} q(j|i,a)u_N(j) \right\} - \frac{\epsilon}{2}.$$

Then we have the following facts.

PROPOSITION 5.2. *Under the conditions in Corollary* 4.3*, the policy f_ϵ obtained by the value iteration algorithm* 5.1 *is ϵ-ASPR-optimal.*

For the policy iteration algorithm 3.1, if we luckily choose an initial policy such that the algorithm 3.1 stops after a *finite* number of iterations, then Proposition 5.1 shows that an ASPR-optimal stationary policy can be computed. Otherwise, since the policy space F may be infinite, the algorithm 3.1 may not stop in any finite number of iterations. In this case, Proposition 5.1 shows that an ASPR-optimal stationary policy can be approximated. On the other hand, Proposition 5.2 implies that under the conditions in Corollary 4.3 an ϵ-ASPR-optimal stationary policy can indeed be computed in a finite number of iterations, where $\epsilon > 0$.

6. Examples. In this section, we illustrate our conditions and show the difference between the ASPR and AER criteria with examples.

Example 6.1 (a controlled birth-death system). Consider a controlled birth-death system in which the state variable denotes the population size at any time $t \geq 0$. There are "natural" birth and death rates denoted by *positive* constants λ and μ, respectively, as well as *nonnegative* emigration and immigration parameters. The two parameters are assumed to be controlled by a decision-maker and denoted by $h_1(i, a_1)$ and $h_2(i, a_2)$, respectively, which may depend on system's state i and decision variables a_1 and a_2 taken by the decision-maker. When the system is at state $i \in S := \{0, 1, \ldots\}$, the decision-maker takes an action $a := (a_1, a_2)$ from a *compact* set $A(i) =: A_1(i) \times A_2(i)$ of available actions, which increases/decreases the emigration parameter $h_1(i, a_1)$ and may incur a cost with rate $c(i, a_1)$, and also increases/decreases the immigration parameter $h_2(i, a_2)$ and gives a reward with rate $\bar{r}(i, a_2)$. Moreover, suppose that the benefit rate caused by a population is represented by $p > 0$. Then the *net* income rate in this system is $r(i, a) := pi + \bar{r}(i, a_2) - c(i, a_1)$ for each $i \in S$ and $a = (a_1, a_2) \in A(i)$. On the other hand, when there is no population in the system (i.e., $i = 0$), it is impossible to decrease/increase the emigration rate,

and so we have $h_1(0,a_1) \equiv 0$ for all $a_1 \in A_1(0)$. Also, in this case (i.e., $i = 0$) we may assume that the decision-maker hopes to increase the immigration rate, and then $h_2(0,a_2) > 0$ for all $a_2 \in A_2(0)$. (This assumption guarantees the irreducibility condition in Assumption A(4).)

We now formulate this system as a continuous-time Markov decision process. The corresponding transition rates $q(j|i,a)$ and reward rates $r(i,a)$ are given as follows.

For $i = 0$ and each $a = (a_1, a_2) \in A(0)$

$$q(1|0,a) = -q(0|0,a) := h_2(0,a_2) > 0,$$

and for $i \geq 1$ and all $a = (a_1, a_2) \in A(i)$

$$q(j|i,a) := \begin{cases} \mu i + h_1(i,a_1) & \text{if } j = i-1, \\ -(\mu+\lambda)i - h_1(i,a_1) - h_2(i,a_2) & \text{if } j = i, \\ \lambda i + h_2(i,a_2) & \text{if } j = i+1, \\ 0 & \text{otherwise;} \end{cases} \tag{6.1}$$

$$r(i,a) := pi + \bar{r}(i,a_2) - c(i,a_1) \quad \text{for } i \in S \text{ and } a = (a_1,a_2) \in A(i). \tag{6.2}$$

We aim to find conditions that ensure the existence of an ASPR-optimal stationary policy. To do this, in the spirit of Assumptions A, B, and C we consider the following conditions:

(E_1) (a) $\mu - \lambda > 0$. (b) Either $\kappa := \mu - \lambda + h_2^* - h_{1*} \leq 0$, or $\mu - \lambda > |h_2^* - h_{1*}|$ when $\kappa > 0$, where $h_2^* := \sup_{a_2 \in A_2(i), i \geq 1} h_2(i,a_2)$, $h_{1*} := \inf_{a_1 \in A_1(i), i \geq 1} h_1(i,a_1)$.

(E_2) For each fixed $i \in S$, the functions $h_1(i,\cdot)$, $h_2(i,\cdot)$, $c(i,\cdot)$, and $\bar{r}(i,\cdot)$ are all continuous.

(E_3) (a) There exist positive constants $L_k (k = 1,2)$ such that $|c(i,a_1)| \leq L_1(i+1)$ and $|\bar{r}(i,a_2)| \leq L_2(i+1)$ for all $i \in S$ and $(a_1,a_2) \in A_1(i) \times A_2(i)$. (b) $\|h_k\| := \sup_{i \in S, a_k \in A_k(i)} |h_k(i,a_k)| < \infty$, for $k = 1,2$.

To further explain Example 6.1, we consider the *special* case of *birth-death processes with controlled immigration*. Consider a pest population in a region which may be isolated to prevent immigration. Let c denote the cost rate when immigration is always prevented, b denote the immigration rate without any control, and action $a \in [0,1]$ denote the *level* of immigration prevented, where c and b are *fixed positive* constants. When the population size is $i \in S := \{0,1,\ldots\}$, an action a from a set $A(i)$ consisting of available actions is taken. Then a cost rate ca is incurred, the immigration rate $(1-a)b$ is permitted, and the evolution of the population depends on birth, death, and immigration with parameters λ, μ, and $(1-a)b$, respectively, where λ and μ are given *positive constants*. Suppose that the damage rate caused by the pest is represented by $p > 0$. Then the reward rate is of the form $r(i,a) := -pi - ca$ for each $i \in S$ and $a \in A(i)$. Obviously, we have $A(i) := [0,1]$ for each $i \geq 1$. However, when there is no pest in the region (i.e., $i = 0$), to guarantee the irreducibility condition in Assumption A(4) we need that $A(0) := [0,\beta]$ with a given $\beta \in (0,1)$. (This, however, can be explained as follows: For the ecological balance of the region, the pest is not permitted to become extinct, and so the immigration rate $(1-\beta)b > 0$ is left.) Using the notation in Example 6.1, for this model we have $h_1 \equiv 0$ and $h_2(i,a_2) = (1-a)b$ with $a_2 := a$ here. Hence, when $\mu - \lambda > b$, the conditions E_1, E_2, and E_3 above are all satisfied.

Under E_1, E_2, and E_3, we obtain the following.

PROPOSITION 6.2. *Under conditions* E_1, E_2, *and* E_3, *the above controlled birth-death system satisfies the Assumptions* A, B, *and* C. *Therefore (by Theorem* 4.1*),*

there exists an ASPR-optimal stationary policy, which can be computed or at least approximated by the policy iteration algorithm 3.1.

Proof. We shall first verify Assumption A. Let $S_n := \{0, 1, \dots, n\}$ for each $n \geq 1$, $w(i) := i + 1$ for all $i \in S$, and

$$\rho := \frac{\mu - \lambda - h_2^* + h_{1*}}{2} = \mu - \lambda - \frac{\kappa}{2} > 0 \quad \text{when } \mu - \lambda > |h_2^* - h_{1*}|.$$

Then Assumptions A(1) and A(2) are obviously true. Moreover, for each $a = (a_1, a_2) \in A(i)$ with $i \geq 1$, by condition E_1 and (6.1), we have

$$\begin{aligned} \sum_{j \in S} q(j|i,a)w(j) &= (\lambda - \mu)(i+1) + \mu - \lambda - h_1(i, a_1) + h_2(i, a_2) \\ &\leq -(\mu - \lambda)w(i) + \kappa \\ &\leq \begin{cases} -(\mu - \lambda)w(i) & \text{when } \kappa \leq 0, \\ -\rho w(i) & \text{when } \kappa > 0 \quad (\text{and so } \rho > 0). \end{cases} \end{aligned} \tag{6.3}$$

In particular, for $i = 0$ and each $a = (a_1, a_2) \in A(0)$, we have

$$\sum_{j \in S} q(j|i,a)w(j) = h_2(0, a_2) \leq -(\mu - \lambda)w(0) + b' = -\rho w(0) + b' - \frac{\kappa}{2}, \tag{6.4}$$

where $b' := \mu - \lambda + \|h_2\| > 0$.

By the inequalities (6.3) and (6.4) we see that Assumption A(3) holds with $c := \mu - \lambda$ and $b := b'$ when $\kappa \leq 0$, or $c := \rho$ and $b := b'$ when $\kappa > 0$. Since $h_2(0, a_2) > 0$ for all $a_2 \in A_2(0)$, by (6.1) we see that Assumption A(4) is true. Hence Assumption A follows.

By E_3 and (6.2), we have $|r(i,a)| \leq pi + L_1(i+1) + L_2(i+1) \leq (p + L_1 + L_2)w(i)$ for all $i \in S$ and $a \in A(i)$, which verifies Assumption B(4). Hence, Assumption B is satisfied because Assumptions B(1), B(2), and B(3) follow from E_2 and the model's description.

Finally, to verify Assumption C we let

$$w_1^*(i) := i^2 + 1, \quad w_2^*(i) := i^4 + 1 \quad \forall i \in S. \tag{6.5}$$

Then

$$w^2(i) \leq M_1^* w_1^*(i), \quad [q(i)w(i)]^2 \leq M_2^* w_2^*(i) \quad \forall i \in S, \tag{6.6}$$

with $M_1^* := 3$ and $M_2^* := 8(\lambda + \mu + \|h_1\| + \|h_2\|)$.

Moreover, for each $i \geq 1$ and $a = (a_1, a_2) \in A(i)$, by (6.1), (6.5), and E_3, we have

$$\begin{aligned} \sum_{j \in S} q(j|i,a)w_1^*(j) &= -2i[\mu i + h_1(i, a_1)] + \mu i + h_1(i, a_1) \\ &\quad + 2i[\lambda i + h_2(i, a_2)] + \lambda i + h_2(i, a_2) \\ &\leq -2(\mu - \lambda)(i^2 + 1) + 3(\mu + \lambda + \|h_1\| + \|h_2\|)i. \end{aligned}$$

Hence, for each $i \geq \frac{3(\mu+\lambda+\|h_1\|+\|h_2\|)}{\mu-\lambda} + 1 =: i_*$, we have

$$\sum_{j \in S} q(j|i,a)w_1^*(j) \leq -(\mu - \lambda)w_1^*(i). \tag{6.7}$$

On the other hand, since $A(i)$ is assumed to be compact for each $i \in S$, by (6.1) and (6.5) we see that $\sum_{j\in S} q(j|i,a)w_1^*(j)$ and $(\mu-\lambda)w_1^*(i)$ are both bounded in $a \in A(i)$ and $i \le i_*$. Thus, from (6.7) there exists a positive constant b_1^* such that

$$\sum_{j\in S} q(j|i,a)w_1^*(j) \le -(\mu-\lambda)w_1^*(i) + b_1^* \quad \forall i \in S \text{ and } a \in A(i). \tag{6.8}$$

Also, for each $i \ge 1$ and $a \in A(i)$, by (6.1) and (6.5) we have

$$\sum_{j\in S} q(j|i,a)w_2^*(j) \le -2(\mu-\lambda)(i^4+1) - (\mu-\lambda)i^4 + c_3 i^3 + c_2 i^2 + c_1 i + c_0, \tag{6.9}$$

where the constants $c_k (k = 0, 1, 2, 3)$ are determined completely by λ, μ, $\|h_1\|$, and $\|h_2\|$. Similarly, by (6.9) and (6.1), there exists a positive constant b_2^* such that

$$\sum_{j\in S} q(j|i,a)w_2^*(j) \le -(\mu-\lambda)w_2^*(i) + b_2^* \quad \forall i \in S \text{ and } a \in A(i), \tag{6.10}$$

which, together with (6.8) and (6.6), verifies Assumption C. □

It should be noted that in Example 6.1 both the reward and transition rates are *unbounded*; see (6.1) and (6.2). Next, we will show that our admissible policy class Π can indeed be chosen to be larger than the usual stationary policy class F.

Example 6.3. In Example 6.1, for each $i \in S$ we take arbitrarily two actions $a^k(i)$ $(k = 1, 2)$ from $A(i)$ which may depend on i, and then define an admissible policy $\tilde{\pi} = (\tilde{\pi}_t)$ as

$$\tilde{\pi}_t(B|i) = \begin{cases} \dfrac{1}{2}e^{-\rho_0 it} & \text{if } B = \{a^1(i)\}, \\ 1 - \dfrac{1}{2}e^{-\rho_0 it} & \text{if } B = \{a^2(i)\}, \\ 0 & \text{otherwise} \end{cases} \tag{6.11}$$

for some fixed constant $\rho_0 > 0$.

Then, by (6.1), (6.11), and (2.2), we see that $\tilde{\pi}$ is in Π but *not* in F. Therefore, we have $\Pi \supset F$, but $\Pi \ne F$. It is also noted that the associated Q-process $p(s, i, t, j, \tilde{\pi})$ is *nonhomogeneous*, and so is the associated continuous-time Markov chain $x(t, \tilde{\pi})$. Moreover, the corresponding reward rates $r(i, \tilde{\pi}_t)$ are *time-dependent* and *unbounded*; see (6.2) and (2.3).

Finally, in the following example we show that *in general* the AER and ASPR criteria are different.

Example 6.4. Let $S := \{1, 2\}$. For some $\hat{\pi} = (\hat{\pi}_t)$, $f \in \Pi$, suppose that for $0 \le t \le 1$,

$$Q(\hat{\pi}_t) := \begin{pmatrix} -1+t & 1-t \\ 2-2t & -2+2t \end{pmatrix} \quad \text{and} \quad Q(f) := \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}. \tag{6.12}$$

Let $t_0 := 1$, and define

$$Q(\tilde{\pi}_t) := \begin{cases} Q(\hat{\pi}_t) & \text{when } 0 \le t \le t_0, \\ Q(f) & \text{when } t \ge t_0. \end{cases} \tag{6.13}$$

By (6.12), (6.13), and Definition 2.2, we see that the associated policy $\tilde{\pi}$ belongs to Π. For reference, we recall that for any $\pi \in \Pi$ the associated regular Q-process $p(s,i,t,j,\pi)$ can be constructed as follows [13, 16, 17]: for $i,j \in S$ and $n \geq 0$, let

$$p_0(s,i,t,j,\pi) := \delta_{ij} e^{-\int_s^t q_i(\pi_y)dy}, \tag{6.14}$$

$$p_{n+1}(s,i,t,j,\pi) := \int_s^t e^{-\int_s^y q_i(\pi_v)dv} \sum_{k\neq i} q(k|i,\pi_y) p_n(y,k,t,j,\pi)dy. \tag{6.15}$$

Then

$$p(s,i,t,j,\pi) = \sum_{n=0}^{\infty} p_n(s,i,t,j,\pi). \tag{6.16}$$

For each $i,j \in S$, by (6.12)–(6.16), $p(0,i,t_0,j,\hat{\pi}) > 0$. Hence, $0 < p(0,i,t_0,2,\hat{\pi}) < 1$. Moreover,

$$p(s,i,t,j,\tilde{\pi}) = \begin{cases} p(s,i,t,j,\hat{\pi}) & \text{when } 0 \leq t \leq t_0, \\ p(s,i,t,j,f) & \text{when } t \geq s \geq t_0. \end{cases} \tag{6.17}$$

Let $r(1,a) = 0$, $r(2,a) = 1$ for all $a \in A(i)$ with $i = 1,2$. Then, by (6.12) and (6.13) we see that states 1 and 2 are absorbing after time t_0. By (6.12), (6.14)–(6.17), we get

$$p(t_0,i,t,i,\tilde{\pi}) = 1 \quad \forall i \in S \text{ and } t \geq t_0. \tag{6.18}$$

Noting that $r(1,\tilde{\pi}_t) = 0$ and $r(2,\tilde{\pi}_t) = 1$ for each $t \geq 0$, by (6.18) and (2.6) we have that for each $i \in S$

$$V_{sp}(\tilde{\pi},i) = 1 \quad \text{for any sample path in } \{x(t,\tilde{\pi}) = 2,\ t \geq t_0\}. \tag{6.19}$$

On the other hand, by the Chapman–Kolmogorov equation and (6.18), we have

$$\begin{aligned} p(0,i,t,2,\tilde{\pi}) &= p(0,i,t_0,1,\tilde{\pi})p(t_0,1,t,2,\tilde{\pi}) + p(0,i,t_0,2,\tilde{\pi})p(t_0,2,t,2,\tilde{\pi}) \\ &= p(0,i,t_0,2,\tilde{\pi}) < 1 \quad \forall t_0 \leq t. \end{aligned} \tag{6.20}$$

Using again $r(1,\tilde{\pi}_t) = 0$ and $r(2,\tilde{\pi}_t) = 1$ for each $t \geq 0$, by (6.20) and (2.7) we get

$$\begin{aligned} \bar{V}(\tilde{\pi},i) &= \limsup_{T\to\infty} \frac{\int_0^T p(0,i,t,2,\tilde{\pi})dt}{T} \\ &= \limsup_{T\to\infty} \frac{\int_{t_0}^T p(0,i,t,2,\tilde{\pi})dt}{T} \\ &= p(0,i,t_0,2,\hat{\pi}) < 1 \quad \forall i \in S, \end{aligned} \tag{6.21}$$

which together with (6.19) and $P_i^{\tilde{\pi}}(\{x(t,\tilde{\pi}) = 2,\ t \geq t_0\}) = p(0,i,t_0,2,\hat{\pi}) > 0$ shows the difference between the ASPR and AER criteria.

7. Concluding remarks. In the previous sections we have studied ASPR optimality for denumerable continuous-time Markov chains determined by possibly unbounded transition rates. Under suitable assumptions we have shown the existence of a solution to the optimality equation and the existence of an ASPR-optimal stationary policy. In addition, we have presented two algorithms to compute, or at least

approximate, the ASPR-optimal stationary policies. Our formulation and approach are sufficiently general and can be used to analyze other important problems, such as the relation among potentials, perturbation analysis, and Markov decision processes in general spaces, as well as minimax control problems. These problems, as far as we can tell, have not been previously studied for continuous-time Markov chains with unbounded transition or reward rates. It should be mentioned that Example 6.4 shows that *in general* the ASPR and AER criteria are different, and it is an interesting and challenging problem to further show the difference between the two criteria under some ergodicity condition. Also, it remains open to show that an ASPR-optimal stationary policy is *not* necessarily canonical. Research on these topics is in progress.

Acknowledgments. The authors are indebted to the anonymous referees for many valuable comments and suggestions that have helped us in improving the presentation.

REFERENCES

[1] E. Altman, *Constrained Markov Decision Processes*, Chapman & Hall/CRC, Boca Raton, FL, 1999.
[2] W.J. Anderson, *Continuous-Time Markov Chains*, Springer-Verlag, New York, 1991.
[3] A. Arapostathis, V.S. Borkar, E. Fernández-Gaucherand, M.K. Ghosh, and S.I. Marcus, *Discrete-time controlled Markov processes with average cost criterion: A survey*, SIAM J. Control Optim., 31 (1993), pp. 282–344.
[4] J. Bather, *Optimal stationary policies for denumerable Markov chains in continuous time*, Adv. in Appl. Probab., 8 (1976), pp. 144–158.
[5] J. Bather, *Optimal decision procedures for finite Markov chains.* II. *Communicating systems*, Adv. in Appl. Probab., 5 (1973), pp. 521–540.
[6] V.S. Borkar, *Topics in Controlled Markov Chains*, Pitman Research Notes in Math. 240, Longman Scientific and Technical, Harlow, UK, 1991.
[7] X.-R. Cao, *The relations among potentials, perturbation analysis, and Markov decision processes*, Discrete Event Dyn. Syst., 8 (1998), pp. 71–87.
[8] X.-R. Cao and H.F. Chen, *Potentials, perturbation realization and sensitivity analysis of Markov processes*, IEEE Trans. Automat. Control, 42 (1997), pp. 1382–1397.
[9] X.-R. Cao and X.P. Guo, *A unified approach to Markov decision problems and performance sensitivity analysis with discounted and average criteria: Multichain cases*, Automatica, 40 (2004), pp. 1749–1759.
[10] R. Cavazos-Cadena and E. Fernández-Gaucherand, *Denumerable controlled Markov chains with average reward criterion: Sample path optimality*, ZOR—Math. Methods Oper. Res., 41 (1995), pp. 89–108.
[11] S.P. Coraluppi and S.I. Marcus, *Risk-sensitive, minimax, and mixed risk-neutral/minimax control of Markov decision processes*, in Stochastic Analysis, Control, Optimization and Applications, Systems Control Found. Appl., Birkhäuser Boston, Boston, 1999, pp. 21–40.
[12] E.B. Dynkin and A.A. Yushkevich, *Controlled Markov Processes*, Springer-Verlag, New York, 1979.
[13] W. Feller, *On the integro-differential equations of purely discontinuous Markoff processes*, Trans. Amer. Math. Soc., 48 (1940), pp. 488–515.
[14] L.G. Gubenko and E.S. Statland, *On discrete time Markov decision processes*, Teor. Verojatnost. i Mat. Statist., 7 (1972), pp. 51–64 (in Russian).
[15] X.P. Guo and P. Shi, *Limiting average criteria for nonstationary Markov decision processes*, SIAM J. Optim., 11 (2001), pp. 1037–1053.
[16] X.P. Guo and O. Hernández-Lerma, *Drift and monotonicity conditions for continuous-time controlled Markov chains with an average criterion*, IEEE Trans. Automat. Control, 48 (2003), pp. 236–245.
[17] X.P. Guo and O. Hernández-Lerma, *Continuous-time controlled Markov chains with discounted rewards*, Acta Appl. Math., 79 (2003), pp. 195–216.
[18] X.P. Guo and K. Liu, *A note on optimality conditions for continuous-time Markov decision processes with average cost criterion*, IEEE Trans. Automat. Control, 46 (2001), pp. 1984–1989.

[19] X.P. Guo and W.P. Zhu, *Denumerable state continuous-time Markov decision processes with unbounded cost and transition rates under average criterion*, ANZIAM J., 43 (2002), pp. 541–557.
[20] M. Haviv and M.L. Puterman, *Bias optimality in controlled queueing systems*, J. Appl. Probab., 35 (1998), pp. 136–150.
[21] O. Hernández-Lerma, *Lectures on Continuous-Time Markov Control Processes*, Aportaciones Matemáticas 3, Sociedad Matematica Mexicana, México City, 1994.
[22] O. Hernández-Lerma, *Adaptive Markov Control Processes*, Springer-Verlag, New York, 1989.
[23] O. Hernández-Lerma and J.B. Lasserre, *Further Topics on Discrete-Time Markov Control Processes*, Springer-Verlag, New York, 1999.
[24] O. Hernández-Lerma, O. Vega-Amaya, and G. Carrasco, *Sample-path optimality and variance-minimization of average cost Markov control processes*, SIAM J. Control Optim., 38 (1999), pp. 79–93.
[25] R.A. Howard, *Dynamic Programming and Markov Processes*, MIT Press, Cambridge, MA, 1960.
[26] P. Kakumanu, *Nondiscounted continuous-time Markov decision processes with countable state space*, SIAM J. Control, 10 (1972), pp. 210–220.
[27] M.E. Lewis and M.L. Puterman, *A note on bias optimality in controlled queueing systems*, J. Appl. Probab., 37 (2000), pp. 300–305.
[28] S.A. Lippman, *On dynamic programming with unbounded rewards*, Management Sci., 21 (1974/75), pp. 1225–1233.
[29] S.A. Lippman, *Applying a new device in the optimization of exponential queuing systems*, Operations Res., 23 (1975), pp. 687–710.
[30] R.B. Lund, S.P. Meyn, and R.L. Tweedie, *Computable exponential convergence rates for stochastically ordered Markov processes*, Ann. Appl. Probab., 6 (1996), pp. 218–237.
[31] M.L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley, New York, 1994.
[32] S.M. Ross, *Applied Probability Models with Optimization Applications*, Holden-Day, San Francisco, 1970.
[33] S.M. Ross, *Introduction to Stochastic Dynamic Programming*, Academic Press, New York, 1983.
[34] S.M. Ross, *Non-discounted denumerable Markovian decision models*, Ann. Math. Statist., 39 (1968), pp. 412–423.
[35] L.I. Sennott, *Stochastic Dynamic Programming and the Control of Queueing Systems*, Wiley, New York, 1999.
[36] R. Serfozo, *Optimal control of random walks, birth and death processes, and queues*, Adv. in Appl. Probab., 13 (1981), pp. 61–83.
[37] R. Serfozo, *An equivalence between continuous and discrete time Markov decision processes*, Oper. Res., 27 (1979), pp. 616–620.
[38] A.F. Veinott, *On finding optimal policies in discrete dynamic programming with no discounting*, Ann. Math. Statist., 37 (1966), pp. 1284–1294.
[39] A.A. Yushkevich and E.A. Feinberg, *On homogeneous Markov model with continuous-time and finite or countable state space*, Theory Probab. Appl., 24 (1979), pp. 156–161.