Hindawi Publishing Corporation
Abstract and Applied Analysis
Volume 2009, Article ID 103723, 17 pages
doi:10.1155/2009/103723

Research Article

Policy Iteration for Continuous-Time Average Reward Markov Decision Processes in Polish Spaces

Quanxin Zhu,[1] Xinsong Yang,[2] and Chuangxia Huang[3]

[1] *Department of Mathematics, Ningbo University, Ningbo 315211, China*
[2] *Department of Mathematics, Honghe University, Mengzi 661100, China*
[3] *The College of Mathematics and Computing Science, Changsha University of Science and Technology, Changsha 410076, China*

Correspondence should be addressed to Quanxin Zhu, zqx22@126.com

Received 24 June 2009; Accepted 9 December 2009

Recommended by Nikolaos Papageorgiou

We study the *policy iteration algorithm* (PIA) for continuous-time jump Markov decision processes in general state and action spaces. The corresponding transition rates are allowed to be *unbounded*, and the reward rates may have *neither upper nor lower bounds*. The criterion that we are concerned with is *expected average reward*. We propose a set of conditions under which we first establish the average reward optimality equation and present the PIA. Then under two *slightly* different sets of conditions we show that the PIA yields the optimal (maximum) reward, an average optimal stationary policy, and a solution to the average reward optimality equation.

1. Introduction

In this paper we study the average reward optimality problem for continuous-time jump Markov decision processes (MDPs) in general state and action spaces. The corresponding transition rates are allowed to be *unbounded*, and the reward rates may have *neither upper nor lower bounds*. Here, the approach to deal with this problem is by means of the well-known policy iteration algorithm (PIA)—also known as Howard's policy improvement algorithm.

As is well known, the PIA was originally introduced by Howard (1960) in [1] for finite MDPs (i.e., the state and action spaces are both finite). By using the monotonicity of the sequence of iterated average rewards, he showed that the PIA converged with a finite number of steps. But, when a state space is not *finite*, there are well-known counterexamples to show that the PIA does not converge even though the action space is compact (see [2–4], e.g.,).

Thus, an interesting problem is to find conditions to ensure that the PIA converges. To do this, extensive literature has been presented; for instance, see [1, 5–14] and the references therein. However, most of those references above are concentrated on the case of discrete-time MDPs; for instance, see [1, 5, 11] for finite discrete-time MDPs, [10, 15] for discrete-time MDPs with a finite state space and a compact action set, [13] for denumerable discrete-time MDPs, and [8, 9, 12] for discrete-time MDPs in Borel spaces. For the case of continuous-time models, to the best of our knowledge, only Guo and Hernández-Lerma [6], Guo and Cao [7], and Zhu [14] have addressed this issue. In [6, 7, 14], the authors established the average reward optimality equation and the existence of average optimal stationary policies. However, the treatments in [6, 7] are restricted to only a denumerable state space. In [14] we used the policy iteration approach to study the average reward optimality problem for the case of continuous-time jump MDPs in general state and action spaces. One of the main contributions in [14] is to prove the existence of the average reward optimality equation and average optimal stationary policies. But the PIA is not stated explicitly in [14], and so the value of the average optimal reward value function and an average optimal stationary policy are also not be computed in [14]. In this paper we further study the average reward optimality problem for such a class of continuous-time jump MDPs in general state and action spaces. Our main objective is to use the PIA to compute or at least approximate (when the PIA takes infinitely many steps to converge) the value of the average optimal reward value function and an average optimal stationary policy. To do this, we first use the so-called "drift" condition, the standard continuity-compactness hypotheses, and the irreducible and uniform exponential *ergodicity* condition to establish the average reward optimality equation and present the PIA. Then under two differently extra conditions we show that the PIA yields the optimal (maximum) reward, an average optimal stationary policy, and a solution to the average reward optimality equation. A key feature of this paper is that the PIA provides an approach to compute or at least approximate (when the PIA takes infinitely many steps to converge) the value of the average optimal reward value function and an average optimal stationary policy.

The remainder of this paper is organized as follows. In Section 2, we introduce the control model and the optimal control problem that we are concerned with. After our optimality conditions and some technical preliminaries as well as the PIA stated in Section 3, we show that the PIA yields the optimal (maximum) reward, an average optimal stationary policy, and a solution to the average reward optimality equation in Section 4. Finally, we conclude in Section 5 with some general remarks.

Notation 1. If X is a Polish space (i.e., a complete and separable metric space), we denote by $\mathcal{B}(X)$ the Borel σ-algebra.

2. The Optimal Control Problem

The material in this section is quite standard (see [14, 16, 17] e.g.,), and we shall state it briefly. The control model that we are interested in is continuous-time jump MDPs with the following form:

$$\{S, (A(x) \subset A,\ x \in S),\ q(\cdot \mid x, a),\ r(x, a)\}, \tag{2.1}$$

where one has the following.

(i) S is a state space and it is supposed to be a Polish space.

(ii) A is an action space, which is also supposed to be a Polish space, and $A(x)$ is a Borel set which denotes the set of available actions at state $x \in S$. The set $K := \{(x, a) : x \in S,\ a \in A(x)\}$ is assumed to be a Borel subset of $S \times A$.

(iii) $q(\cdot \mid x, a)$ denotes the *transition rates*, and they are supposed to satisfy the following properties: for each $(x, a) \in K$ and $D \in \mathcal{B}(\mathrm{S})$,

(Q_1) $D \mapsto q(D \mid x, a)$ is a signed measure on $\mathcal{B}(S)$, and $(x, a) \mapsto q(D \mid x, a)$ is Borel measurable on K;

(Q_2) $0 \le q(D \mid x, a) < \infty$, for all $x \notin D \in \mathcal{B}(\mathrm{S})$;

(Q_3) $q(S \mid x, a) = 0,\ 0 \le -q(x \mid x, a) < \infty$;

(Q_4) $q(x) := \sup_{a \in A(x)}(-q(x \mid x, a)) < \infty$, for all $x \in S$.

It should be noted that the property (Q_3) shows that the model is *conservative*, and the property (Q_4) implies that the model is *stable*.

(iv) $r(x, a)$ denotes the *reward rate* and it is assumed to be measurable on K. (As $r(x, a)$ is allowed to take positive and negative values; it can also be interpreted as a *cost rate*.)

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

Let Π_m be the family of function $\pi_t(B \mid x)$ such that

(i) for each $x \in S$ and $t \ge 0$, $B \to \pi_t(B \mid x)$ is a probability measure on $\mathcal{B}(A(x))$,

(ii) for each $x \in S$ and $B \in \mathcal{B}(A(x))$, $t \to \pi_t(B \mid x)$ is a Borel measurable function on $[0, \infty)$.

Definition 2.1. A family $\pi = (\pi_t, t \ge 0) \in \Pi_m$ is said to be a *randomized Markov policy*. In particular, if there exists a measurable function f on S with $f(x) \in A(x)$ for all $x \in S$, such that $\pi_t(\{f(x)\} \mid x) \equiv 1$ for all $t \ge 0$ and $x \in S$, then π is called a (deterministic) *stationary policy* and it is identified with f. The set of all stationary policies is denoted by F.

For each $\pi = (\pi_t, t \ge 0) \in \Pi_m$, we define the associated transition rates $q(D \mid x, \pi_t)$ and the reward rates $r(x, \pi_t)$, respectively, as follows.

For each $x \in S$, $D \in \mathcal{B}(S)$ and $t \ge 0$,

$$\begin{aligned} q(D \mid x, \pi_t) &:= \int_{A(x)} q(D \mid x, a)\pi_t(da \mid x), \\ r(x, \pi_t) &:= \int_{A(x)} r(x, a)\pi_t(da \mid x). \end{aligned} \tag{2.2}$$

In particular, we will write $q(D \mid x, \pi_t)$ and $r(x, \pi_t)$ as $q(D \mid x, f)$ and $r(x, f)$, respectively, when $\pi := f \in F$.

Definition 2.2. A randomized Markov policy is said to be *admissible* if $q(D \mid x, \pi_t)$ is continuous in $t \ge 0$, for all $D \in \mathcal{B}(S)$ and $x \in S$.

The family of all such policies is denoted by Π. Obviously, $\Pi \supseteq F$ and so that Π is nonempty. Moreover, for each $\pi \in \Pi$, Lemma 2.1 in [16] ensures that there exists

a Q-process—that is, a possibly substochastic and nonhomogeneous transition function $P^{\pi}(s,x,t,D)$ with transition rates $q(D \mid x,\pi_t)$. As is well known, such a Q-process is not necessarily regular; that is, we might have $P^{\pi}(s,x,t,S) < 1$ for some state $x \in S$ and $t \geq s \geq 0$. To ensure the regularity of a Q-process, we shall use the following so-called "drift" condition, which is taken from [14, 16–18].

Assumption A. There exist a (measurable) function $w_1 \geq 1$ on S and constants $b_1 \geq 0$, $c_1 > 0$, $M_1 > 0$ and $M > 0$ such that

(1) $\int_S w_1(y)q(dy \mid x,a) \leq -c_1 w_1(x) + b_1$ for all $(x,a) \in K$;

(2) $q(x) \leq M_1 w_1(x)$ for all $x \in S$, with $q(x)$ as in (Q_4);

(3) $|r(x,a)| \leq M w_1(x)$ for all $(x,a) \in K$.

Remark 2.1 in [16] gives a discussion of Assumption A. In fact, Assumption A(1) is similar to conditions in the previous literature (see [19, equation (2.4)] e.g.,), and it is together with Assumption A(3) used to ensure the finiteness of the average expected reward criterion (2.5) below. In particular, Assumption A(2) is not required when the transition rate is uniformly bounded, that is, $\sup_{x\in S} q(x) < \infty$.

For each initial state $x \in S$ at time $s \geq 0$ and $\pi \in \Pi$, we denote by $P^{\pi}_{s,x}$ and $E^{\pi}_{s,x}$ the probability measure determined by $P^{\pi}(s,x,t,D)$ and the corresponding expectation operator, respectively. Thus, for each $\pi \in \Pi$ by [20, pages 107–109] there exists a Borel measure Markov process $\{x^{\pi}_t\}$ (we shall denote $\{x^{\pi}_t\}$ by $\{x_t\}$ for simplicity when there is no risk of confusion) with value in S and the transition function $P^{\pi}(s,x,t,D)$, which is completely determined by the transition rates $q(D \mid x,\pi_t)$. In particular, if $s = 0$, we write $E^{\pi}_{0,x}$ and $P^{\pi}_{0,x}$ as E^{π}_x and P^{π}_x, respectively.

If Assumption A holds, then from [17, Lemma 3.1] we have the following facts.

Lemma 2.3. *Suppose that Assumption A holds. Then the following statements hold.*

(a) *For each $x \in S$, $\pi \in \Pi$ and $t \geq 0$,*

$$E^{\pi}_x[w_1(x_t)] \leq e^{-c_1 t} w_1(x) + \frac{b_1}{c_1}, \tag{2.3}$$

where the function w_1 and constants b_1 and c_1 are as in Assumption A.

(b) *For each $u \in B_{w_1}(S)$, $x \in S$ and $\pi \in \Pi$,*

$$\lim_{t\to\infty} \frac{E^{\pi}_x[u(x_t)]}{t} = 0. \tag{2.4}$$

For each $x \in S$ and $\pi \in \Pi$, the *expected average reward* $V(x,\pi)$ as well as the corresponding optimal reward value functions $V^*(x)$ are defined as

$$V(x,\pi) := \liminf_{T\to\infty} \frac{\int_0^T [E^{\pi}_x r(x_t,\pi_t)]dt}{T}, \qquad V^*(x) := \sup_{\pi\in\Pi} V(x,\pi). \tag{2.5}$$

As a consequence of Assumption A(3) and Lemma 2.3(a), the expected average reward $V(x,\pi)$ is well defined.

Definition 2.4. A policy $\pi^* \in \Pi$ is said to be *average optimal* if $V(x,\pi^*) = V^*(x)$ for all $x \in S$.

The main goal of this paper is to give conditions for ensuring that the policy iteration algorithm converges.

3. Optimality Conditions and Preliminaries

In this section we state conditions for ensuring that the policy iteration algorithm (PIA) converges and give some preliminary lemmas that are needed to prove our main results.

To guarantee that the PIA converges, we need to establish the average reward optimality equation. To do this, in addition to Assumption A, we also need two more assumptions. The first one is the following so-called standard continuity-compactness hypotheses, which is taken from [14, 16–18]. Moreover, it is similar to the version for discrete-time MDPs; see, for instance, [3, 8, 21–23] and their references. In particular, Assumption B(3) is not required when the transition rate is uniformly bounded, since it is only used to ensure the applying of the *Dynkin formula*.

Assumption B. For each $x \in S$,

(1) $A(x)$ is compact;

(2) $r(x,a)$ is continuous 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 $u := w_1$ as in Assumption A;

(3) there exist a nonnegative measurable function w_2 on S, and constants $b_2 \geq 0$, $c_2 > 0$ and $M_2 > 0$ such that

$$q(x)w_1(x) \leq M_2 w_2(x), \qquad \int_S w_2(y)q(dy \mid x,a) \leq c_2 w_2(x) + b_2 \tag{3.1}$$

for all $(x,a) \in K$.

The second one is the irreducible and uniform exponential *ergodicity* condition. To state this condition, we need to introduce the concept of the weighted norm used in [8, 14, 22]. For the function $w_1 \geq 1$ in Assumption A, we define the weighted supremum norm $\|\cdot\|_{w_1}$ for real-valued functions u on S by

$$\|u\|_{w_1} := \sup_{x \in S}\left[w_1(x)^{-1}|u(x)|\right] \tag{3.2}$$

and the Banach space

$$B_{w_1}(S) := \left\{u : \|u\|_{w_1} < \infty\right\}. \tag{3.3}$$

Definition 3.1. For each $f \in F$, the Markov process $\{x_t\}$, with transition rates $q(\cdot \mid x, f)$, is said to be *uniform w_1-exponentially ergodic* if there exists an invariant probability measure μ_f on S

such that

$$\sup_{f\in F}\left|E_x^f[u(x_t)]-\mu_f(u)\right|\le \mathrm{Re}^{-\rho t}\|u\|_{w_1}w_1(x) \tag{3.4}$$

for all $x\in S$, $u\in B_{w_1}(S)$ and $t\ge 0$, where the positive constants R and ρ do not depend on f, and where $\mu_f(u):=\int_S u(y)\mu_f(dy)$.

Assumption C. For each $f\in F$, the Markov process $\{x_t\}$, with transition rates $q(\cdot\mid x,f)$, is uniform w_1-exponentially *ergodic* and λ-*irreducible*, where λ is a nontrivial σ-finite measure on $\mathcal{B}(S)$ independent of f.

Remark 3.2. (a) Assumption C is taken from [14] and it is used to establish the average reward optimality equation. (b) Assumption C is similar to the uniform w_1-exponentially ergodic hypothesis for discrete-time MDPs; see [8, 22], for instance. (c) Some sufficient conditions as well as examples in [6, 16, 19] are given to verify Assumption C. (d) Under Assumptions A, B, and C, for each $f\in F$, the Markov process $\{x_t\}$, with the transition rate $q(\cdot\mid x,f)$, has a unique invariant probability measure μ_f such that

$$\int_S \mu_f(dx)q(D\mid x,f)=0\quad\text{for each } D\in\mathcal{B}(S). \tag{3.5}$$

(e) As in [9], for any given stationary policy $f\in F$, we shall also consider two functions in $B_{w_1}(S)$ to be equivalent and do not distinguish between equivalent functions, if they are equal μ_f-almost everywhere (a.e.). In particular, if $u(x)=0$ μ_f-a.e. holds for all $x\in S$, then the function u will be taken to be identically zero.

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

Lemma 3.3. *Suppose that Assumptions A, B, and C hold, and let $f\in F$ be any stationary policy. Then one has the following facts.*

(a) *For each $x\in S$, the function*

$$h_f(x):=\int_0^\infty\left[E_x^f(r(x_t,f))-g(f)\right]dt \tag{3.6}$$

belongs to $B_{w_1}(S)$, where $g(f):=\int_S r(y,f)\mu_f(dy)$ and w_1 is as in Assumption A.

(b) *$(g(f),h_f)$ satisfies the Poisson equation*

$$g(f)=r(x,f)+\int_S h_f(y)q(dy\mid x,f)\qquad\forall x\in S, \tag{3.7}$$

for which the μ_f-expectation of h_f is zero, that is,

$$\mu_f(h_f):=\int_S h_f(y)\mu_f(dy)=0. \tag{3.8}$$

(c) *For all* $x \in S$, $|V(x,f)| \le Mb_1/c_1$.

(d) *For all* $x \in S$, $|g(f)| = |V(x,f)| \le Mb_1/c_1$.

Proof. Obviously, the proofs of parts (a) and (b) are from [14, Lemma 3.2]. We now prove (c). In fact, from the definition of $V(x,f)$ in (2.5), Assumption A(3), and Lemma 2.3(a) we have

$$\left|V(x,f)\right| \le \liminf_{T\to\infty} \frac{\int_0^T M\left[e^{-c_1 t} w_1(x) + b_1/c_1\right] dt}{T} = \frac{Mb_1}{c_1}, \tag{3.9}$$

which gives (c). Finally, we verify part (d). Obviously, by Assumption A(3) and Assumption C we can easily obtain $g(f) = V(x,f)$ for all $x \in S$, which together with part (c) yields the desired result. □

The next result establishes the *average reward optimality equation*. For the proof, see [14, Theorem 4.1].

Theorem 3.4. *Under Assumptions A, B, and C, the following statements hold.*

(a) *There exist a unique constant* g^*, *a function* $h^* \in B_{w_1}(S)$, *and a stationary policy* $f^* \in F$ *satisfying the average reward optimality equation*

$$g^* = \max_{a\in A(x)} \left\{ r(x,a) + \int_S h^*(y) q(dy \mid x,a) \right\} \tag{3.10}$$

$$= r(x,f^*) + \int_S h^*(y) q(dy \mid x, f^*) \quad \forall x \in S. \tag{3.11}$$

(b) $g^* = \sup_{\pi\in\Pi} V(x,\pi)$ *for all* $x \in S$.

(c) *Any stationary policy* $f \in F$ *realizing the maximum of* (3.10) *is average optimal, and so* f^* *in* (3.11) *is average optimal.*

Then, under Assumptions A, B, and C we shall present the PIA that we are concerned with. To do this, we first give the following definition.

For any real-valued function u on S, we define the dynamic programming operator T as follows:

$$Tu(x) := \max_{a\in A(x)} \left\{ r(x,a) + \int_S u(y) q(dy \mid x,a) \right\} \quad \forall x \in S. \tag{3.12}$$

Algorithm A (policy iteration).

Step 1 (initialization). Take $n = 0$ and choose a stationary policy $f_n \in F$.

Step 2 (policy evaluation). Find a constant $g(f_n)$ and a real-valued function h_{f_n} on Ssatisfying the Poisson equation (3.7), that is,

$$g(f_n) = r(x,f_n) + \int_S h_{f_n}(y) q(dy \mid x, f_n) \quad \forall x \in S. \tag{3.13}$$

Obviously, by (3.12) and (3.13) we have

$$g(f_n) \le Th_{f_n}(x) = \max_{a\in A(x)}\left\{r(x,a) + \int_S h_{f_n}(y)q(dy \mid x,a)\right\} \quad \forall x \in S. \tag{3.14}$$

Step 3 (policy improvement). Set $f_{n+1}(x) := f_n(x)$ for all $x \in S$ for which

$$r(x,f_n) + \int_S h_{f_n}(y)q(dy \mid x,f_n) = Th_{f_n}(x); \tag{3.15}$$

otherwise (i.e., when (3.15) does not hold), choose $f_{n+1}(x) \in A(x)$ such that

$$r(x,f_{n+1}) + \int_S h_{f_n}(y)q(dy \mid x,f_{n+1}) = Th_{f_n}(x). \tag{3.16}$$

Step 4. If f_{n+1} satisfies (3.15) for all $x \in S$, then stop (because, from Proposition 4.1 below, f_{n+1} is average optimal); otherwise, replace f_n with f_{n+1} and go back to Step 2.

Definition 3.5. The policy iteration Algorithm A is said to *converge* if the sequence $\{g(f_n)\}$ converges to the average optimal reward value function in (2.5), that is,

$$\lim_{n\to\infty} g(f_n) = V^*(x) = g^* \quad \forall x \in S, \tag{3.17}$$

where g^* is as in Theorem 3.4.

Obviously, under Assumptions A, B, and C from Proposition 4.1 we see that the sequence $\{g(f_n)\}$ is nondecreasing; that is, $g(f_n) \le g(f_{n+1})$ holds for all $n \ge 1$. On the other hand, by Lemma 3.3(d) we see that $\{g(f_n)\}$ is bounded. Therefore, there exists a constant g such that

$$\lim_{n\to\infty} g(f_n) = g. \tag{3.18}$$

Noting that, in general, we have $g \le g^*$. In order to ensure that the policy iteration Algorithm A converges, that is, $g = g^*$, in addition to Assumptions A, B, and C, we need an additional condition (Assumption D (or D′) below).

Assumption D. There exist a subsequence $\{h_{f_m}\}$ of $\{h_{f_n}\}$ and a measurable function h on S such that

$$\lim_{m\to\infty} h_{f_m}(x) = h(x) \quad \forall x \in S. \tag{3.19}$$

Remark 3.6. (a) Assumption D is the same as the hypothesis H1 in [9], and Remark 4.6 in [9] gives a detailed discussion of Assumption D. (b) In particular, Assumption D trivially holds when the state space S is a *countable* set (with the discrete topology). (c) When the state space S is not *countable*, if the sequence $\{h_{f_n}\}$ is equicontinuous, Assumption D also holds.

Assumption D′. There exists a stationary policy $f^* \in F$ such that

$$\lim_{n\to\infty} f_n(x) = f^*(x) \quad \forall x \in S. \tag{3.20}$$

Remark 3.7. Assumption D′ is the same as the hypothesis H2 in [9]. Obviously, Assumption D′ trivially holds when the state space S is a *countable* set (with the discrete topology) and $A(x)$ is compact for all $x \in S$.

Finally, we present a lemma (Lemma 3.8) to conclude this section, which is needed to prove our Theorem 4.2. For a proof, see [24, Proposition 12.2], for instance.

Lemma 3.8. *Suppose that $A(x)$ is compact for all $x \in S$, and let $\{f_n\}$ be a stationary policy sequence in F. Then there exists a stationary policy $f \in F$ such that $f(x) \in A(x)$ is an accumulation point of $\{f_n(x)\}$ for each $x \in S$.*

4. Main Results

In this section we will present our main results, Theorems 4.2-4.3. Before stating them, we first give the following proposition, which is needed to prove our main results.

Proposition 4.1. *Suppose that Assumptions A, B, and C hold, and let $f \in F$ be an arbitrary stationary policy. If any policy $\overline{f} \in F$ such that*

$$Th_f(x) = r\left(x, \overline{f}\right) + \int_S h_f(y) q\left(dy \mid x, \overline{f}\right) \quad \forall x \in S, \tag{4.1}$$

then (a)

$$g(f) \le g\left(\overline{f}\right); \tag{4.2}$$

(b) *if $g(f) = g(\overline{f})$, then*

$$h_f(\cdot) = h_{\overline{f}}(\cdot) + k \quad \text{for some constant } k; \tag{4.3}$$

(c) *if f is average optimal, then*

$$h_f(\cdot) = h^*(\cdot) + k' \quad \text{for some constant } k', \tag{4.4}$$

where h^* is as in Theorem 3.4;

(d) *if $g(f) = g(\overline{f})$, then $(g(f), h_f)$ satisfies the average reward optimality equation* (3.10)*, and so f is average optimal.*

Proof. (a) Combining (3.7) and (4.1) we have

$$g(f) \le r\left(x, \overline{f}\right) + \int_S h_f(y) q\left(dy \mid x, \overline{f}\right) \quad \forall x \in S. \tag{4.5}$$

Obviously, taking the integration on both sides of (4.5) with respect to $\mu_{\overline{f}}$ and by Remark 3.2(d) we obtain the desired result.

(b) If $g(f) = g(\overline{f})$, we may rewrite the Poisson equation for $\overline{f}$ as

$$g(f) = r\left(x, \overline{f}\right) + \int_S h_{\overline{f}}(y) q\left(dy \mid x, \overline{f}\right) \quad \forall x \in S. \tag{4.6}$$

Then, combining (4.5) and (4.6) we obtain

$$\int_S \left[h_f(y) - h_{\overline{f}}(y)\right] q\left(dy \mid x, \overline{f}\right) \ge 0 \quad \forall x \in S. \tag{4.7}$$

Thus, from (4.7) and using the Dynkin formula we get

$$E_x^{\overline{f}}\left[h_f(x_t) - h_{\overline{f}}(x_t)\right] \ge h_f(x) - h_{\overline{f}}(x) \quad \forall x \in S. \tag{4.8}$$

Letting $t \to \infty$ in (4.8) and by Assumption C we have

$$\mu_{\overline{f}}\left(h_f - h_{\overline{f}}\right) \ge h_f(x) - h_{\overline{f}}(x) \quad \forall x \in S. \tag{4.9}$$

Now take $k := \sup_{x \in S}[h_f(x) - h_{\overline{f}}(x)]$. Then take the supremum over $x \in S$ in (4.9) to obtain

$$k \ge \mu_{\overline{f}}\left(h_f - h_{\overline{f}}\right) \ge k, \tag{4.10}$$

and so

$$\mu_{\overline{f}}\left(h_f - h_{\overline{f}}\right) = k, \tag{4.11}$$

which implies

$$h_f(\cdot) = h_{\overline{f}}(\cdot) + k \quad \mu_{\overline{f}}\text{-a.e.} \tag{4.12}$$

Hence, from Remark 3.2(e) and (4.12) we obtain (4.3).

(c) Since f is average optimal, by Definition 2.4 and Theorem 3.4(b) we have

$$g(f) = g^* = \sup_{\pi \in \Pi} V(x, \pi) \quad \forall x \in S. \tag{4.13}$$

Hence, the Poisson equation (3.7) for f becomes

$$g^* = r(x,f) + \int_S h_f(y)q(dy \mid x,f) \quad \forall x \in S. \tag{4.14}$$

On the other hand, by (3.10) we obtain

$$\begin{aligned} g^* &= \max_{a\in A(x)}\left\{r(x,a) + \int_S h^*(y)q(dy \mid x,a)\right\} \\ &\geq r(x,f) + \int_S h^*(y)q(dy \mid x,f) \quad \forall x \in S, \end{aligned} \tag{4.15}$$

which together with (4.14) gives

$$\int_S [h_f(y) - h^*(y)]q(dy \mid x,f) \geq 0 \quad \forall x \in S. \tag{4.16}$$

Thus, as in the proof of part (b), from (4.16) we see that (4.4) holds with $k' := \sup_{x\in S}[h_f(x) - h^*(x)]$.

(d) By (3.7), (4.1), (4.3), and Q_3 we have

$$\begin{aligned} g(f) &\leq Th_f(x) \\ &= r\left(x,\overline{f}\right) + \int_S \left[h_{\overline{f}}(y) + k\right]q\left(dy \mid x,\overline{f}\right) \\ &= r\left(x,\overline{f}\right) + \int_S h_{\overline{f}}(y)q\left(dy \mid x,\overline{f}\right) \\ &= g\left(\overline{f}\right) = g(f) \quad \forall x \in S, \end{aligned} \tag{4.17}$$

which gives

$$g(f) = Th_f(x) \quad \forall x \in S, \tag{4.18}$$

that is,

$$g(f) = \max_{a\in A(x)}\left\{r(x,a) + \int_S h_f(y)q(dy \mid x,a)\right\} \quad \forall x \in S. \tag{4.19}$$

Thus, as in the proof of Theorem 4.1 in [14], from Lemma 2.3(b), (3.7), and (4.19) we show that f is average optimal, that is, $g(f) = g^*$. Hence, we may rewrite (4.19) as

$$g^* = \max_{a\in A(x)}\left\{r(x,a) + \int_S h_f(y)q(dy \mid x,a)\right\} \quad \forall x \in S. \tag{4.20}$$

Thus, from (4.20) and part (c) we obtain the desired conclusion. □

Theorem 4.2. *Suppose that Assumptions A, B, C, and D hold, then the policy iteration Algorithm A converges.*

Proof. From Lemma 3.3(a) we see that the function h_{f_n} in (3.13) belongs to $B_{w_1}(S)$, and so the function h in (3.19) also belongs to $B_{w_1}(S)$. Now let $\{h_{f_n}\}$ be as in Assumption D, and let $\{h_{f_m}\}$ be the corresponding subsequence of $\{h_{f_n}\}$. Then by Assumption D we have

$$\lim_{m\to\infty} h_{f_m}(x) = h(x) \quad \forall x \in S. \tag{4.21}$$

Moreover, from Lemma 3.8 there is a stationary policy $f \in F$ such that $f(x) \in A(x)$ is an accumulation point of $\{f_m(x)\}$ for each $x \in S$; that is, for each $x \in S$ there exists a subsequence $\{m_i\}$ (depending on the state x) such that

$$\lim_{i\to\infty} f_{m_i}(x) = f(x) \quad \forall x \in S. \tag{4.22}$$

Also, by (3.13) we get

$$g(f_{m_i}) = r(x, f_{m_i}) + \int_S h_{f_{m_i}}(y) q(dy \mid x, f_{m_i}) \quad \forall x \in S. \tag{4.23}$$

On the other hand, take any real-valued measurable function m on S such that $m(x) > q(x) \geq 0$ for all $x \in S$. Then, for each $x \in S$ and $a \in A(x)$, by the properties (Q_1)–(Q_3) we can define $P(\cdot \mid x, a)$ as follows:

$$P(D \mid x, a) := \frac{q(D \mid x, a)}{m(x)} + I_D(x) \quad \forall D \in \mathcal{B}(S). \tag{4.24}$$

Obviously, $P(\cdot \mid x, a)$ is a probability measure on S. Thus, combining (4.23) and (4.24) we have

$$g(f_{m_i}) = r(x, f_{m_i}) + \int_S h_{f_{m_i}}(y) P(dy \mid x, f_{m_i}) \quad \forall x \in S. \tag{4.25}$$

Letting $i \to \infty$ in (4.25), then by (3.18), (4.21), and (4.22) as well as the "extension of Fatou's lemma" 8.3.7 in [8] we obtain

$$g = r(x, f) + \int_S h(y) q(dy \mid x, f) \quad \forall x \in S. \tag{4.26}$$

To complete the proof of Theorem 4.2, by Proposition 4.1(d) we only need to prove that g, h, and f satisfy the average reward optimality equation (3.10) and (3.11), that is,

$$g = Th(x) = r(x, f) + \int_S h(y) q(dy \mid x, f) \quad \forall x \in S. \tag{4.27}$$

Obviously, from (4.26), and the definition of T in (3.12) we obtain

$$g \le Th(x) \quad \forall x \in S. \tag{4.28}$$

The rest is to prove the reverse inequality, that is,

$$g \ge Th(x) \quad \forall x \in S. \tag{4.29}$$

Obviously, by (3.19) we have

$$\lim_{i\to\infty}\left[h_{f_{m_i}}(x) - h_{f_{m_{i-1}}}(x)\right] = 0 \quad \forall x \in S. \tag{4.30}$$

Moreover, from Lemma 3.3(a) again we see that there exists a constant k such that

$$\left\|h_{f_n}\right\|_{w_1} \le k \quad \forall n \ge 1, \tag{4.31}$$

which gives

$$\left\|h_{f_{m_i}} - h_{f_{m_{i-1}}}\right\|_{w_1} \le \left\|h_{f_{m_i}}\right\|_{w_1} + \left\|h_{f_{m_{i-1}}}\right\|_{w_1} \le 2k. \tag{4.32}$$

Thus, by (4.24), (4.31), (4.32) and the "extension of Fatou's lemma" 8.3.7 in [8] we obtain

$$\lim_{i\to\infty}\int_S \left[h_{f_{m_i}}(y) - h_{f_{m_{i-1}}}(y)\right] P(dy \mid x, f_{m_i}) = 0 \quad \forall x \in S, \tag{4.33}$$

which implies

$$\lim_{i\to\infty}\int_S \left[h_{f_{m_i}}(y) - h_{f_{m_{i-1}}}(y)\right] q(dy \mid x, f_{m_i}) = 0 \quad \forall x \in S. \tag{4.34}$$

Also, from (3.7), (3.16), and the definition of T in (3.12) we get

$$\begin{aligned} g(f_{m_i}) &= r(x, f_{m_i}) + \int_S h_{f_{m_i}}(y) q(dy \mid x, f_{m_i}) \\ &= Th_{f_{m_{i-1}}}(x) + \int_S \left[h_{f_{m_i}}(y) - h_{f_{m_{i-1}}}(y)\right] q(dy \mid x, f_{m_i}) \\ &\ge r(x,a) + \int_S h_{f_{m_{i-1}}}(y) q(dy \mid x, a) \\ &\quad + \int_S \left[h_{f_{m_i}}(y) - h_{f_{m_{i-1}}}(y)\right] q(dy \mid x, f_{m_i}) \quad \forall x \in S,\ a \in A(x). \end{aligned} \tag{4.35}$$

Letting $i \to \infty$ in (4.35), then by (3.18), (4.21), (4.22), (4.34), and the "extension of Fatou's lemma 8.3.7" in [8] we obtain

$$g \geq r(x,a) + \int_S h(y)q(dy \mid x,a) \quad \forall x \in S,\ a \in A(x), \tag{4.36}$$

which gives

$$g \geq \max_{a \in A(x)} \left\{ r(x,a) + \int_S h(y)q(dy \mid x,a) \right\} = Th(x) \quad \forall x \in S. \tag{4.37}$$

This completes the proof of Theorem 4.2. □

Theorem 4.3. *Suppose that Assumptions A, B, C, and D′ hold, then the policy iteration Algorithm A converges.*

Proof. To prove Theorem 4.3, from the proof of Theorem 4.2 we only need to verify that (4.26) and (4.27) hold true for f^* as in Assumption D′ and some function h in $B_{w_1}(S)$. To do this, we first define two functions h, h' in $B_{w_1}(S)$ as follows:

$$h(x) := \limsup_{n \to \infty} h_{f_n}(x), \quad h'(x) := \liminf_{n \to \infty} h_{f_n}(x) \quad \forall x \in S. \tag{4.38}$$

Then by (3.7) we get

$$g(f_n) = r(x,f_n) + \int_S h_{f_n}(y)q(dy \mid x,f_n) \quad \forall x \in S, \tag{4.39}$$

which together with (4.24) yields

$$g(f_n) = r(x,f_n) + \int_S h_{f_n}(y)P(dy \mid x,f_n) \quad \forall x \in S. \tag{4.40}$$

Applying the "extension of Fatou's Lemma 8.3.7" in [8] and letting $n \to \infty$ in (4.40), then by (3.18), (4.38) and Assumption D′ we obtain

$$\begin{aligned} g &\leq r(x,f^*) + \int_S h(y)P(dyx,f^*) \quad \forall x \in S, \\ g &\geq r(x,f^*) + \int_S h'(y)P(dy \mid x,f^*) \quad \forall x \in S, \end{aligned} \tag{4.41}$$

which implies

$$g \le r(x, f^*) + \int_S h(y) q(dy \mid x, f^*) \quad \forall x \in S, \tag{4.42}$$

$$g \ge r(x, f^*) + \int_S h'(y) q(dy \mid x, f^*) \quad \forall x \in S. \tag{4.43}$$

Thus, combining (4.42) and (4.43) we get

$$\int_S [h(y) - h'(y)] q(dy \mid x, f^*) \ge 0 \quad \forall x \in S. \tag{4.44}$$

Then, from the proof of Proposition 4.1(b) and (4.44) we have

$$h(\cdot) = h'(\cdot) + k'' \quad \text{for some constant } k'', \tag{4.45}$$

which together with (4.42), (4.43), and the definition of T in (3.12) gives

$$g = r(x, f^*) + \int_S h(y) q(dy \mid x, f^*) \le Th(x) \quad \forall x \in S. \tag{4.46}$$

The remainder is to prove the reverse inequality, that is,

$$g \ge Th(x) = \max_{a \in A(x)} \left\{ r(x, a) + \int_S h(y) q(dy \mid x, a) \right\} \quad \forall x \in S. \tag{4.47}$$

Obviously, by (3.16) and (4.24) we get

$$\frac{r(x, f_{n+1})}{m(x)} + \int_S h_{f_n}(y) P(dy \mid x, f_{n+1}) \ge \frac{r(x, a)}{m(x)} + \int_S h_{f_n}(y) P(dy \mid x, a) \quad \forall x \in S,\ a \in A(x). \tag{4.48}$$

Then, letting $n \to \infty$ in (4.48), by (4.38), Assumption D′, and the "extension of Fatou's Lemma 8.3.7" in [8], we obtain

$$\frac{r(x, f^*)}{m(x)} + \int_S h(y) P(dy \mid x, f^*) \ge \frac{r(x, a)}{m(x)} + \int_S h(y) P(dy \mid x, a) \quad \forall x \in S,\ a \in A(x), \tag{4.49}$$

which implies

$$r(x, f^*) + \int_S h(y) q(dy \mid x, f^*) \ge r(x, a) + \int_S h(y) q(dy \mid x, a) \quad \forall x \in S,\ a \in A(x), \tag{4.50}$$

and so

$$r(x,f^*)+\int_S h(y)q(dy\mid x,f^*)\geq \max_{a\in A(x)}\left\{r(x,a)+\int_S h(y)q(dy\mid x,a)\right\}\quad \forall x\in S. \tag{4.51}$$

Thus, combining (4.46) and (4.51) we see that (4.47) holds. And so Theorem 4.3 follows. □

5. Concluding Remarks

In the previous sections we have studied the policy iteration algorithm (PIA) for average reward continuous-time jump MDPs in Polish spaces. Under two *slightly* different sets of conditions we have shown that the PIA yields the optimal (maximum) reward, an average optimal stationary policy, and a solution to the average reward optimality equation. It should be mentioned that the approach presented here is different from the policy iteration approach used in [14] because the PIA in this paper provides an approach to compute or at least approximate (when the PIA takes infinitely many steps to converge) the value of the average optimal reward value function and an average optimal stationary policy.

Acknowledgments

The author would like to thank the editor and anonymous referees for their good comments and valuable suggestions, which have helped us to improve the paper. This work was jointly supported by the National Natural Science Foundation of China (10801056), the Natural Science Foundation of Ningbo (201001A6011005) the Scientific Research Fund of Zhejiang Provincial Education Department, K.C. Wong Magna Fund in Ningbo University, the Natural Science Foundation of Yunnan Provincial Education Department (07Y10085), the Natural Science Foundation of Yunnan Provincial (2008CD186), the Foundation of Chinese Society for Electrical Engineering (2008).

References

[1] R. A. Howard, *Dynamic Programming and Markov Processes*, The Technology Press of M.I.T., Cambridge, Mass, USA, 1960.
[2] R. Dekker, "Counter examples for compact action Markov decision chains with average reward criteria," *Communications in Statistics*, vol. 3, no. 3, pp. 357–368, 1987.
[3] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons, New York, NY, USA, 1994.
[4] P. J. Schweitzer, "On undiscounted Markovian decision processes with compact action spaces," *RAIRO—Operations Research*, vol. 19, no. 1, pp. 71–86, 1985.
[5] E. V. Denardo and B. L. Fox, "Multichain Markov renewal programs," *SIAM Journal on Applied Mathematics*, vol. 16, pp. 468–487, 1968.
[6] X. P. Guo and O. Hernández-Lerma, "Drift and monotonicity conditions for continuous-time controlled Markov chains with an average criterion," *IEEE Transactions on Automatic Control*, vol. 48, no. 2, pp. 236–245, 2003.
[7] X. P. Guo and X. R. Cao, "Optimal control of ergodic continuous-time Markov chains with average sample-path rewards," *SIAM Journal on Control and Optimization*, vol. 44, no. 1, pp. 29–48, 2005.
[8] O. Hernández-Lerma and J. B. Lasserre, *Further Topics on Discrete-Time Markov Control Processes*, vol. 42 of *Applications of Mathematics*, Springer, New York, NY, USA, 1999.

[9] O. Hernández-Lerma and J. B. Lasserre, "Policy iteration for average cost Markov control processes on Borel spaces," *Acta Applicandae Mathematicae*, vol. 47, no. 2, pp. 125–154, 1997.
[10] A. Hordijk and M. L. Puterman, "On the convergence of policy iteration in finite state undiscounted Markov decision processes: the unichain case," *Mathematics of Operations Research*, vol. 12, no. 1, pp. 163–176, 1987.
[11] J. B. Lasserre, "A new policy iteration scheme for Markov decision processes using Schweitzer's formula," *Journal of Applied Probability*, vol. 31, no. 1, pp. 268–273, 1994.
[12] S. P. Meyn, "The policy iteration algorithm for average reward Markov decision processes with general state space," *IEEE Transactions on Automatic Control*, vol. 42, no. 12, pp. 1663–1680, 1997.
[13] M. S. Santos and J. Rust, "Convergence properties of policy iteration," *SIAM Journal on Control and Optimization*, vol. 42, no. 6, pp. 2094–2115, 2004.
[14] Q. X. Zhu, "Average optimality for continuous-time Markov decision processes with a policy iteration approach," *Journal of Mathematical Analysis and Applications*, vol. 339, no. 1, pp. 691–704, 2008.
[15] A. Y. Golubin, "A note on the convergence of policy iteration in Markov decision processes with compact action spaces," *Mathematics of Operations Research*, vol. 28, no. 1, pp. 194–200, 2003.
[16] X. P. Guo and U. Rieder, "Average optimality for continuous-time Markov decision processes in Polish spaces," *The Annals of Applied Probability*, vol. 16, no. 2, pp. 730–756, 2006.
[17] Q. X. Zhu, "Average optimality inequality for continuous-time Markov decision processes in Polish spaces," *Mathematical Methods of Operations Research*, vol. 66, no. 2, pp. 299–313, 2007.
[18] Q. X. Zhu and T. Prieto-Rumeau, "Bias and overtaking optimality for continuous-time jump Markov decision processes in Polish spaces," *Journal of Applied Probability*, vol. 45, no. 2, pp. 417–429, 2008.
[19] R. B. Lund, S. P. Meyn, and R. L. Tweedie, "Computable exponential convergence rates for stochastically ordered Markov processes," *The Annals of Applied Probability*, vol. 6, no. 1, pp. 218–237, 1996.
[20] I. I. Gīhman and A. V. Skorohod, *Controlled Stochastic Processes*, Springer, New York, NY, USA, 1979.
[21] Q. X. Zhu and X. P. Guo, "Markov decision processes with variance minimization: a new condition and approach," *Stochastic Analysis and Applications*, vol. 25, no. 3, pp. 577–592, 2007.
[22] Q. X. Zhu and X. P. Guo, "Another set of conditions for Markov decision processes with average sample-path costs," *Journal of Mathematical Analysis and Applications*, vol. 322, no. 2, pp. 1199–1214, 2006.
[23] Q. X. Zhu and X. P. Guo, "Another set of conditions for strong $n(n = -1, 0)$ discount optimality in Markov decision processes," *Stochastic Analysis and Applications*, vol. 23, no. 5, pp. 953–974, 2005.
[24] M. Schäl, "Conditions for optimality in dynamic programming and for the limit of n-stage optimal policies to be optimal," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 32, no. 3, pp. 179–196, 1975.