

A new strong optimality criterion for nonstationary Markov decision processes*

Xianping Guo¹, Peng Shi², Weiping Zhu³

¹ Department of Mathematics, Zhongshan University, Guangzhou 510275, P. R. China
(e-mail: mcsgxp@zsu.edu.cn)

² Land Operations Division, Defence Science and Technology Organisation, PO Box 1500,
Salisbury 5108 SA, Australia (e-mail: peng.shi@dsto.defence.gov.au)

³ Department of Computer Science and Electrical Engineering, The University of Queensland,
St. Lucia 4072, QLD, Australia

Manuscript received: December 1999/Final version received: May 2000

Abstract. This paper deals with a new optimality criterion consisting of the usual three average criteria and the canonical triplet (totally so-called strong average-canonical optimality criterion) and introduces the concept of a strong average-canonical policy for nonstationary Markov decision processes, which is an extension of the canonical policies of Hernández-Lerma and Lasserre [16] (pages: 77) for the stationary Markov controlled processes. For the case of possibly non-uniformly bounded rewards and denumerable state space, we first construct, under some conditions, a solution to the optimality equations (OEs), and then prove that the Markov policies obtained from the OEs are not only optimal for the three average criteria but also optimal for all finite horizon criteria with a sequence of additional functions as their terminal rewards (i.e. strong average-canonical optimal). Also, some properties of optimal policies and optimal average value convergence are discussed. Moreover, the error bound in average reward between a rolling horizon policy and a strong average-canonical optimal policy is provided, and then a rolling horizon algorithm for computing strong average $\varepsilon(>0)$ -optimal Markov policies is given.

Key words: Nonstationary Markov decision processes, optimality equations, strong average-canonical optimal policies

* This work has been supported partially by the Natural Science Foundation of China (No. 19901038), by the Natural Science Foundation of Guangdong Province, by the Foundation of Hong Kong and Zhongshan University Advanced Research Center, by the Center for Industrial and Applicable Mathematics, The University of South Australia, and by the University of Queensland under Grant No. 8/UQNSRG025G.

1 Introduction

As is well known, the finite and infinite horizon Markov decision processes (MDPs, for short) have been an interesting subject, and many authors have devoted to this topic. Among this research line, the so-called long-run “average reward” and finite-horizon models are one of the most commonly adopted versions, in which the widely employed average criteria are *the average expected criterion* (see Arapostathis, Borker, Gaucherand, Ghosh and Markus [2], Derman [6], Dynkin and Yushkevich [7], Filar and Vrieze [10], Hernández-Lerma [15], Hernández-Lerma and Lasserre [16] and Puterman [21], etc.), *the expected average criterion* (see Bieth [4], Blackwell [5], Feinberg and Park [8], and Filar, Krass and Ross [9], etc.), and *the sample path average criterion* (see Rolando and Emanuel [22], Ross and Varadarajan [23] and [24], etc.). However, the majority of the researchers have centered their emphasis on the stationary case, that is, both rewards and transition probabilities are time free. Indeed, in real world, both of the above two elements may be changed with time. Hence, it is more interesting and practical to investigate the situation of both rewards and transition probabilities being dependent on time (i.e., non-stationary case). For nonstationary Markov decision processes (NMDPs, for short) with expected totally reward criterion, some excellent results are available, for instance, [7] and [17]. In relation to the NMDPs some work has been done. Now we briefly summarize the main results on NMDPs. For the case of finite state and action spaces, Hopp, Bean and Smith [18] show that an accumulation point of a sequence of finite horizon optimal policies is optimal for the average expected criterion, and, Alden and Smith [1] provide an error bound in average expected cost between a rolling horizon policy and an average expected optimal policy. Bean, Smith and Lasserre [3] extend the results for the above finite state model to denumerable state case. Park, Bean and Smith [20] prove that the optimal finite horizon average values converge to the infinite horizon optimal average expected value in denumerable state case, and also show that, by an example, even if the NMDPs satisfies the weak ergodicity widely used in the discussions on the average expected criterion, the stationary MDPs deduced from the NMDPs by the traditional transformation may not. By using optimality equations, Hou and Guo [19, 11] prove the existence of average expected optimal Markov policies, and Guo [12, 13] discuss the properties of the average expected optimal policies and the average variance criterion. It should be noted that all the rewards in [18, 1, 3, 20, 19, 11, 12, 13] are assumed to be uniformly bounded. For the case of non-uniformly rewards, Guo, et al. [14] discuss the average expected criterion. It is well known that the above three average criteria are the preferred criteria in many applications. However, they are obviously under-selective, that is, two policies π and π' may have the same average reward value, but π may outperform π' for all finite-horizon criteria and the errors in finite-horizon rewards between policies π and π' may be unbounded in stages. Therefore, the concept of a canonical policy, which is not only optimal for the average expected criterion, but also optimal for any finite horizon problems with the additive terminal rewards, has been introduced, and the existence of canonical policies are studied (see [7, 2, 16], etc.). But the treatment in [7, 2, 16] is restricted to the stationary case. For the nonstationary case, the existence of canonical policies seems not discussed yet. The aim of this paper is to do some work on this area.

In this paper, we shall consider a new criterion consisting of the above three

average criteria and the canonical triplet (totally so-called strong average-canonical optimality criterion) and introduce the concept of a strong average-canonical policy for NMDPs, which is an extension of the canonical policies of Hernández-Lerma and Lasserre [16] (pages: 77) and stronger than each of the above three average criteria. For the case of possibly non-uniformly bounded rewards and denumerable state space, we first construct, under some conditions, a solution to the optimality equations (OEs), and then show that the Markov policies obtained from the OEs are not only optimal for the three average criteria but also optimal for all finite horizon criteria with a sequence of additional functions as their terminal rewards (i.e. strong average-canonical optimal). Also, some properties of the optimal policies and optimal average expected value convergence are discussed. Especially, an example is given in which all conditions in this paper are satisfied, but some conditions in [18, 1, 3, 20, 19, 11, 12, 13] fail to hold. Moreover, the error bounds in the three average rewards between a rolling horizon policy and a strong average-canonical optimal policy are provided, and then a rolling horizon algorithm for computing strong average $\varepsilon (> 0)$ -optimal Markov policies is presented.

The paper continues as follows: in Section 2, the notation and the definitions are introduced. The average optimality equations for the underlying model are established, and the existence of solutions to the OE's is also presented in Section 3. The existence of strong average-canonical optimal Markov policies and their properties are given in Section 4. Section 5 provides an algorithm for computing strong average-canonical $\varepsilon > 0$ -optimal Markov policies based the error bounds.

2 Model, notation and definitions

The model considered in this paper is a six-element tuple $\{S_n, A_n, (A_n(i)|i \in S_n, n \geq 0), (P_n), (r_n), W\}$ consisting of

- (a) a denumerable space S_n , the state space at time n , with Borel σ -algebra $\mathcal{B}(S_n)$ generated by all subsets of S_n ;
- (b) a Borel space A_n , the action space at time n , with Borel σ -algebra $\mathcal{B}(A_n)$;
- (c) a family $\{A_n(i)|i \in S_n, n \geq 0\}$ of nonempty measurable subsets of A_n , where $A_n(i)$ denotes the set of feasible actions when the system is in state $i \in S_n$ at stage n . Let

$$K_n := \{(i, a)|i \in S_n, a \in A_n(i)\}, \quad \forall n \geq 0;$$

- (d) stochastic kernel P_n on S_{n+1} given K_n , that is the transition probability of the system from stage n to stage $n + 1$;
- (e) a measurable function $r_n(n \geq 0) : K_n \rightarrow R$, that is the n stage reward functions;
- (f) a strong average reward criterion W (see in sequel).

For each $n = 0, 1, \dots$, we define the space H_n of admissible histories up to time n as $H_0 := S_0$ and $H_n := K_0 \times K_1 \times \dots \times K_{n-1} \times S_n, \forall n \geq 1$. A generic element $h_n \in H_n$, which is an admissible n -history, is a vector of the form $h_n = (i_0, a_0, \dots, i_{n-1}, a_{n-1}, i_n)$ with $(i_t, a_t) \in K_t, \forall t = 0, 1, \dots, n - 1$, and $i_n \in S_n$. Of course, for each $n \geq 0, H_n$ is a subspace of $(S_n \times A_n)^n \times S_{n+1}$.

A randomized policy is a sequence $\pi = \{\pi_n, n = 0, 1, \dots\}$, where stochastic kernel π_n on the action space A_n given H_n satisfies

$$\pi_n(A_n(i_n)|h_n) = 1, \quad \forall h_n \in H_n, \quad n = 0, 1, \dots \tag{2.1}$$

The set of all randomized policies is denoted by Π . A randomized policy $\pi = \{\pi_n, n = 0, 1, \dots\} \in \Pi$ is called randomized Markov one, if $\pi_n(\cdot|h_n) = \pi_n(\cdot|i_n)$, $\forall h_n \in H_n, n \geq 0$. The set of all randomized Markov policies is denoted by Π_m . Let $F_n (n \geq 0)$ be the set of all functions $f_n : S_n \rightarrow A_n$ satisfying $f_n(i) \in A_n(i)$, $\forall i \in S_n$. A randomized Markov policy $\pi = \{\pi_n, n \geq 0\}$ is called a Markov policy if for every $n \geq 0$ there exists a $f_n \in F_n$ such that $\pi_n(f_n(i)|i) = 1, \forall i \in S_n, n \geq 0$. Clearly, a Markov policy can be represented as a sequence of mappings $\{f_n\}$ such that $f_n \in F_n$ for all $n \geq 0$. The set of all Markov policies is denoted by Π_m^d . Obviously, $\Pi_m^d \subset \Pi_m \subset \Pi$.

For any $\pi \in \Pi$ and $i \in S_0$, by the result of Ionescu-Tulcea (see [16] pages 179 and 16), there exists a unique probability measure P_π^i on $((S_n \times A_n)^\infty, (\mathcal{B}(S_n) \times \mathcal{B}(A_n))^\infty)$ such that $P_\pi^i(H_\infty) = 1$, and, for all $i \in S_0, j \in S_{n+1}, n = 0, 1, \dots$

$$P_\pi^i(X_0 = i) = 1, \tag{2.2}$$

$$P_\pi^i(X_{n+1} = j|X_0, A_0, \dots, X_n, A_n) = P_n(j|X_n, A_n), \tag{2.3}$$

where X_n and A_n are state and action variables at stage n , respectively. The expectation operator with respect to P_π^i is denoted by E_π^i . Since $P_\pi^i(H_\infty) = 1$, the average expected criterion \bar{V} , the expected average criterion \bar{U} , the sample path average criterion V_S and the finite-horizon criterion J_N are defined well respectively as follows: for $\pi \in \Pi, i \in S_0$ and $N \geq 1$, let

$$\bar{V}(\pi, i) := \limsup_{N \rightarrow \infty} \frac{\sum_{n=0}^{N-1} E_\pi^i r_n(X_n, A_n)}{N}, \tag{2.4}$$

$$\bar{U}(\pi, i) := E_\pi^i \limsup_{N \rightarrow \infty} \frac{\sum_{n=0}^{N-1} r_n(X_n, A_n)}{N}, \tag{2.5}$$

$$V_S(\pi, i) := \limsup_{N \rightarrow \infty} \frac{\sum_{n=0}^{N-1} r_n(X_n, A_n)}{N}, \tag{2.6}$$

$$J_N(\pi, i) := E_\pi^i \left[\sum_{n=0}^{N-1} r_n(X_n, A_n) \right]. \tag{2.7}$$

Definition 2.1. Let $\{g_n\}$ be a real number sequence, $\{u_n\}$ be a real-value function sequence and $\pi \in \Pi_m^d$. Then (g_n, u_n, π) is said to be a canonical triplet if, for every $i \in S$ and $N \geq 1$,

$$J_N(\pi^*, i, u_N) := J_N(\pi^*, i) + E_{\pi^*}^i u_N(X_N) = \sum_{n=0}^{N-1} g_n + u_0(i) \tag{2.8}$$

$$\geq J_N(\pi, i, u_N) := J_N(\pi, i) + E_\pi^i u_N(X_N). \tag{2.9}$$

A policy $\pi \in \Pi$ is said to be canonical if it enters into some canonical triplet.

Obviously, a canonical policy is an optimal for all finite-horizon problems with u_N as the terminal rewards.

Definition 2.2. For any $\varepsilon \geq 0$, a policy $\pi^* \in \Pi$ is called \bar{V} - ε -optimal, if $\bar{V}(\pi^*, i) \geq \bar{V}(\pi, i) - \varepsilon$, $\forall i \in S_0$ and $\pi \in \Pi$. A \bar{V} -0-optimal policy is called \bar{V} -optimal. Similarly, we can define \bar{U} - ε -optimal policies and \bar{U} -optimal policies.

Definition 2.3. For any $\varepsilon \geq 0$, a policy $\pi^* \in \Pi$ is called V_s - ε -optimal, if there exists a constant ρ such that

$$P_{\pi^*}^i \{V_s(\pi^*, i) \geq \rho - \varepsilon\} = 1, \quad \forall i \in S_0, \quad (2.10)$$

$$P_{\pi}^i \{V_s(\pi, i) \leq \rho\} = 1, \quad \forall i \in S_0, \pi \in \Pi. \quad (2.11)$$

A V_s -0-optimal policy is called V_s -optimal.

Definition 2.4. For any $\varepsilon \geq 0$, a policy $\pi^* \in \Pi$ is called strong average ε -optimal, if π^* is w - ε -optimal for any $w \in \{\bar{V}, \bar{U}, V_s\}$. A strong average 0-optimal policy is called strong average optimal.

Definition 2.5. A policy $\pi^* \in \Pi$ is called strong average-canonical optimal, if π^* is strong average optimal and canonical.

It can be easily seen that strong average-canonical optimality is stronger than any one of average expected optimality, expected average optimality, sample path average optimality, and is the generalization of the canonical criterion for stationary MDPs (see [16] (Page: 77), etc.).

3 Optimality equations

In this section, we shall establish the OE's for NMDP's and provide some conditions under which a solution to the OE's can be constructed.

Let $M(S_n)$ denote the set of all real-value functions on S_n .

Definition 3.1. If there exist a real number sequence $\{g_n\}$ and a sequence of functions $u_n \in M(S_n)$ such that

$$g_n + u_n(i) = \sup_{a \in A_n(i)} \left\{ r_n(i, a) + \sum_{j \in S_{n+1}} P_n(j|i, a) u_{n+1}(j) \right\},$$

$$\forall i \in S_n, n \geq 0, \quad (3.1)$$

then we call the functional equations (3.1) the OE's for our model and both the sequences $\{g_n\}$ and $\{u_n\}$ a solution to the OE's (3.1).

In order to construct a solution to the OE's (3.1), we need the following conditions:

Assumption 3.1. For any $n \geq 0$, there exists a measure δ_n on S_{n+1} satisfying

$$(i) \quad \delta_n(S_{n+1}) \neq 1, \quad \text{and} \quad \delta_n(j) \geq (\text{or } \leq) P_n(j|i, a), \quad \forall i \in S_n, a \in A_n(i); \tag{3.2}$$

$$(ii) \quad \|r_n\| + \sum_{t=0}^{\infty} |(1 - \delta_n(S_{n+1})) \cdots (1 - \delta_{n+t}(S_{n+t+1}))| \|r_{n+t+1}\| := R_n < \infty, \tag{3.3}$$

where $\|r_n\| := \sup_{i \in S_n, a \in A_n(i)} |r_n(i, a)|$.

Remark 3.1. It should be noted that Assumption 3.1 is the extension of the mirror conditions in Dynkin and Yushkevich [7], and the conditions in Park et al. [20], Bean et al. [3] and in Alden and Smith [1]. Also, Assumption 3.1 is weaker than those in Park et al. [20] and in Alden and Smith [1].

If $\sum_{j \in S_{n+1}} \sup_{i \in S_n, a \in A_n(i)} P_n(j|i, a) \neq 1$ or $\sum_{j \in S_{n+1}} \inf_{i \in S_n, a \in A_n(i)} P_n(j|i, a) \neq 1$ for all $n \geq 0$, then we take a fixed measure δ_n on S_{n+1} defined as $\delta_n(j) := \sup_{i \in S_n, a \in A_n(i)} P_n(j|i, a), j \in S_{n+1}$ or $\delta_n(j) = \inf_{i \in S_n, a \in A_n(i)} P_n(j|i, a), j \in S_{n+1}$.

In order to construct a solution to the OEs, we let, for all $n \geq 0, i \in S_n, a \in A_n(i)$ and $\pi \in \Pi$,

$$\beta_0 = 1, \quad \beta_n := (1 - \delta_0(S_1)) \cdots (1 - \delta_{n-1}(S_n)); \tag{3.4}$$

$$\bar{P}_n(\cdot|i, a) := [P_n(\cdot|i, a) - \delta_n(\cdot)] / (1 - \delta_n(S_{n+1})); \tag{3.5}$$

$$G_N(\pi, i) := \bar{E}_\pi^i \left[\sum_{n=N}^{\infty} \beta_n r_n(i, a) | X_N = i \right]; \tag{3.6}$$

$$G_N^*(i) := \sup_{\pi \in \Pi} G_N(\pi, i), \tag{3.7}$$

where, \bar{E}_π^i denotes the corresponding expected operator for the above new transition probabilities.

Now, we present our results on the solution to the OE's.

Theorem 3.1. If Assumption 3.1 holds, then a solution $\{g_n\}$ and $\{u_n\}$ of the OEs (3.1) satisfying $\|u_n\| \leq |R_n| \forall n \geq 0$, can be constructed as follows:

$$u_n(i) := G_n^*(i) / \beta_n \quad \forall n \geq 0; \tag{3.8}$$

$$g_n := \sum_{j \in S_{n+1}} u_{n+1}(j) \delta_n(j) \quad \forall n \geq 0. \tag{3.9}$$

Proof. Since $\sum_{n=0}^{\infty} |\beta_n| \|r_n\| < \infty$, by Theorem 9.2 in [17], we have

$$\begin{aligned}
 G_n^*(i) &= \sup_{a \in A_n(i)} \left\{ (1 - \delta_0(S_1)) \cdots (1 - \delta_{n-1}(S_n)) r_n(i, a) + \sum_{j \in S_{n+1}} \bar{P}_n(j|i, a) G_{n+1}^*(j) \right\} \\
 &= \sup_{a \in A_n(i)} \left\{ \beta_n r_n(i, a) + \sum_{j \in S_{n+1}} [P_n(j|i, a) - \delta_n(j)] / (1 - \delta_n(S_{n+1})) G_{n+1}^*(j) \right\} \\
 &= \sup_{a \in A_n(i)} \left\{ \beta_n r_n(i, a) + \beta_n \sum_{j \in S_{n+1}} (P_n(j|i, a) - \delta_n(j)) [G_{n+1}^*(j) / \beta_{n+1}] \right\}. \quad (3.10)
 \end{aligned}$$

From (3.10), (3.9) and (3.8), we have

$$g_n + u_n(i) = \sup_{a \in A_n(i)} \left\{ r_n(i, a) + \sum_{j \in S_{n+1}} P_n(j|i, a) u_{n+1}(j) \right\}, \quad (3.11)$$

which yields (3.1). From (3.8) and (3.6), we can derive $\|u_n\| \leq |R_n|, \forall n \geq 0$. Hence the proof of this theorem is completed. $\nabla \nabla \nabla$

Assumption 3.2. (i) For every $n \geq 0$ and $i \in S_n, A_n(i)$ is compact;

(ii) For every $n \geq 0$ and $i \in S_n, the reward function $r_n(i, a)$ is upper semi-continuous in a on $A_n(i)$;$

(iii) For every $n \geq 0, i \in S, the function $v'(i, a) := \sum_{j \in S_{n+1}} P_n(j|i, a)v(j)$ is upper semi-continuous in a on $A_n(i)$ for every $v \in M(S_{n+1})$.$

Theorem 3.2. If Assumptions 3.1 and 3.2 hold, we have

- (i) there exist a number sequence $\{g_n\}$ and a function sequence $\{u_n\}$ satisfying (3.1), $\|u_n\| \leq \|R_n\|, \forall n \geq 0$;
- (ii) there exists a Markov policy $\pi^* = \{f_n^*\} \in \Pi_m^d$ such that for all $i \in S_n$ and $n \geq 0$,

$$r_n(i, f_n^*(i)) + \sum_{j \in S_{n+1}} P_n(j|i, f_n^*(i)) u_{n+1}(j) = g_n + u_n(i). \quad (3.12)$$

Proof. By Theorem 3.1, there exist sequences $\{g_n\}$ and $\{u_n\}$ satisfying (3.1) and $\|u_n\| \leq \|R_n\| < \infty, \forall n \geq 0$. Hence, by Assumption 3.2, part (i) is valid. Part (ii) follows from part (i) and Lemma 5.6 in [17]. $\nabla \nabla \nabla$

4 Strong average-canonical optimality

In this section we will derive the existence of strong average ε -optimal and strong average-canonical optimal Markov policies from the OE's (3.1), then analyze the properties of optimal policies and prove the convergence of the optimal average expected value. The approach employed here is rather different from those used in [18, 3, 20, 1]. The martingale theory is used to develop our main results.

Theorem 4.1. *If $\{g_n\}$ and $\{u_n\}$ are a solution to the OE's (3.1) and satisfy $\sum_{n=1}^{\infty} \frac{\|u_n\|^2}{n} < \infty$, then, for all $i \in S_0, \pi \in \Pi$ and $N \geq 1$,*

$$(a) \quad J_N(\pi, i, u_N) \leq \sum_{n=0}^{N-1} g_n + u_0(i); \tag{4.1}$$

$$(b) \quad \bar{V}(\pi, i) \leq \rho^*; \tag{4.2}$$

$$(c) \quad \bar{U}(\pi, i) \leq \rho^*; \tag{4.3}$$

$$(d) \quad P_{\pi}^i\{V_s(\pi, i) \leq \rho^*\} = 1, \tag{4.4}$$

where, $\rho^* := \limsup_{N \rightarrow \infty} \frac{g_0 + g_1 + \dots + g_{N-1}}{N}$.

Proof. For any $i \in S_0, \pi \in \Pi$ and $n \geq 0$, since $\sum_{n=1}^{\infty} \frac{\|u_n\|^2}{n^2} < \infty$, we have

$$\lim_{n \rightarrow \infty} \frac{\|u_n\|}{n} = 0. \tag{4.5}$$

Therefore, $|E_{\pi}^i u_n(X_n)| \leq \|u_n\| < \infty \forall n \geq 0$. From (2.3) and (3.1), we obtain

$$\begin{aligned} & E_{\pi}^i[u_{n+1}(X_{n+1})|X_0, \Delta_0, \dots, X_n, \Delta_n] \\ &= \sum_{j \in S_{n+1}} u_{n+1}(j) P_n(j|X_n, \Delta_n) \\ &= r_n(X_n, \Delta_n) + \sum_{j \in S_{n+1}} u_{n+1}(j) P_n(j|X_n, \Delta_n) - r_n(X_n, \Delta_n) \\ &\leq g_n + u_n(X_n) - r_n(X_n, \Delta_n). \end{aligned} \tag{4.6}$$

By taking expectation operator E_{π}^i on the both sides of (4.6), we have

$$E_{\pi}^i[u_{n+1}(X_{n+1})] \leq g_n + E_{\pi}^i[u_n(X_n)] - E_{\pi}^i[r_n(X_n, \Delta_n)]. \tag{4.7}$$

By induction, for every $N \geq 1$, (4.7) gives

$$E_{\pi}^i[u_N(X_N)] + \sum_{n=0}^{N-1} E_{\pi}^i[r_n(X_n, \Delta_n)] \leq \sum_{n=0}^{N-1} g_n + u_0(i), \tag{4.8}$$

which yields part (a).

From (4.5) and (4.8), one has

$$\bar{V}(\pi, i) \leq \limsup_{N \rightarrow \infty} \frac{\sum_{n=0}^{N-1} g_n}{N} = \rho^*,$$

which completes the proof of part (b).

To prove parts (c) and (d), for any $i \in S_n$ and $a \in A_n(i)$, $n \geq 0$, let

$$D_n(i, a) := r_n(i, a) + \sum_{j \in S_{n+1}} u_{n+1}(j) P_n(j|i, a) - g_n - u_n(i). \quad (4.9)$$

Since $\{g_n\}$ and $\{u_n\}$ are a solution to the OE's (3.1), we have

$$D_n(i, a) \leq 0, \quad \forall i \in S_n, a \in A_n(i) \quad \text{and} \quad n \geq 0. \quad (4.10)$$

For any $h = (i_0, a_0, \dots, i_n, a_n, \dots) \in (S_n \times A_n)^\infty$ and $n \geq 0$, let

$$Z_n(h) := \begin{cases} r_n(i_n, a_n) + u_{n+1}(i_{n+1}) - g_n - u_n(i_n) - Z_n(i_n, a_n), & \text{if } h_n \in H_n; \\ 0, & \text{if } h_n \notin H_n. \end{cases}$$

From (2.1), we have

$$\begin{aligned} E_\pi^i[Z_n|X_0, A_0, \dots, X_n, A_n] \\ = E_\pi^i[u_{n+1}(X_{n+1})|X_0, A_0, \dots, X_n, A_n] - \sum_{j \in S} u_{n+1}(j) P_n(j|X_n, A_n) = 0. \end{aligned}$$

Hence, $\{\sum_{n=0}^{N-1} Z_n, \sigma(X_0, A_0, \dots, X_{N-1}, A_{N-1})\}$ is a martingale. By $E_\pi^i u_n^2(X_n) \leq \|u_n\|^2 < \infty$, we can also derive $\{\sum_{n=0}^{N-1} Z_n\}$ is square-integrable. Obviously, $\{N, \sigma(X_0, A_0, \dots, X_{N-1}, A_{N-1})\}$ is predictable increasing. By Theorem 7.5.4 in [25], we have

$$\lim_{N \rightarrow \infty} \frac{\sum_{n=0}^{N-1} Z_n}{N} = 0, \quad a.e. -P_\pi^i, \quad (4.11)$$

which gives us

$$\begin{aligned} \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} [r_n(X_n, A_n) + u_{n+1}(X_{n+1}) - u_n(X_n) - g_n - D_n(X_n, A_n)] = 0, \\ a.e. -P_\pi^i. \end{aligned} \quad (4.12)$$

Since $(-D_n(X_n, A_n)) \geq 0$, a.e. $-P_\pi^i$ and $\lim_{N \rightarrow \infty} \frac{1}{N} u_N(X_N) = 0$, a.e. $-P_\pi^i$, from (4.12), we have

$$\begin{aligned} 0 &\geq \limsup_{N \rightarrow \infty} \left[\frac{1}{N} \sum_{n=0}^{N-1} r_n(X_n, A_n) - \frac{1}{N} \sum_{n=0}^{N-1} g_n + \frac{1}{N} u_N(X_N) \right] \\ &\geq \limsup_{N \rightarrow \infty} \left[\frac{1}{N} \sum_{n=0}^{N-1} r_n(X_n, A_n) \right] + \liminf_{N \rightarrow \infty} - \left[\frac{1}{N} \sum_{n=0}^{N-1} g_n \right], \quad a.e. -P_\pi^i, \end{aligned}$$

which means

$$\limsup_{N \rightarrow \infty} \left[\frac{1}{N} \sum_{n=0}^{N-1} r_n(X_n, A_n) \right] \leq \limsup_{N \rightarrow \infty} \left[\frac{1}{N} \sum_{n=0}^{N-1} g_n \right] = \rho^*, \quad a.e. -P_\pi^i. \quad (4.13)$$

Therefore, part (c) is valid.

To prove part (d), we take expectation operator on both sides of (4.13), then

$$\bar{U}(\pi, i) \leq \rho^*. \quad (4.14)$$

Note that π and i are arbitrary, the desired result can be obtained from (4.14). ▽▽▽

In order to prove the existence of the strong average-canonical Markov policies, and then to discuss the properties of ε -optimal policies, for a given sequence of numbers $\varepsilon_n \geq 0$ and $i \in S_n$, we let

$$A_n^{\varepsilon_n}(i) := \left\{ a \in A_n(i) : r_n(i, a) + \sum_{j \in S_{n+1}} P_n(j|i, a)u_{n+1}(j) \geq g_n + u_n(i) - \varepsilon_n \right\}$$

and

$$\Pi^*(\varepsilon_n) := \{ \pi \in \Pi : P_\pi^i(A_n^{\varepsilon_n}(i_n)|h_n) = 1 \forall h_n = (i_0, a_0, \dots, i_n) \in H_n, n \geq 0 \}.$$

If $\varepsilon_n = 0(\varepsilon), \forall n \geq 0$, then we write $\Pi^*(\varepsilon_n)$ as $\Pi^*(0)(\Pi^*(\varepsilon))$.

Theorem 4.2. (i) *If Assumptions 3.1 holds and all $\varepsilon_n > 0$, then there exists a Markov policy $\pi \in \Pi^*(\varepsilon_n)$.*

(ii) *If Assumptions 3.1 and 3.2 hold, then there exists a Markov policy $\pi \in \Pi^*(0)$.*

Proof. Under Assumption 3.1, by Theorem 3.1 we can obtain that $\|u_n\| \leq R_n < \infty, \forall n \geq 0$. Hence, for any $i \in S_n, n \geq 0$, there exists a $f_n(i) \in A_n^{\varepsilon_n}(i)$, thus, the corresponding Markov policy $\pi := (f_0, \dots, f_n, \dots) \in \Pi^*(\varepsilon)$. This means part (i) is valid. By Theorems 3.1 and 3.2, part (ii) can be carried out. ▽▽▽

Theorem 4.3. *Suppose that $\{g_n\}$ and $\{u_n\}$ are a solution to the OE's (3.1) and satisfy $\sum_{n=1}^\infty \frac{\|u_n\|^2}{n} < \infty$.*

(i) *If a policy $\pi \in \Pi^*(\varepsilon_n)$, then π is strong average ε -optimal, here $\varepsilon := \limsup_{N \rightarrow \infty} \frac{\sum_{n=0}^{N-1} \varepsilon_n}{N}$,*

(ii) *If a policy $\pi^* \in \Pi^*(0)$, then*

- (a) $P_{\pi^*}^i \{V_s(\pi^*, i) = \bar{V}(\pi^*, i) = \bar{U}(\pi^*, i) = \rho^*\} = 1, \forall i \in S_0$, and
- (b) *the policy π^* is strong average-canonical optimal.*

Proof. (i) Since, for all $i \in S_n, a \in A_n^{\varepsilon_n}(i)$ and $n \geq 0$,

$$r_n(i, a) + \sum_{j \in S_{n+1}} P_n(j|i, a)u_{n+1}(j) \geq u_n(i) + g_n - \varepsilon_n, \quad (4.15)$$

and, $\pi \in \Pi^*(\varepsilon)$, i.e. $P_\pi^i(A_n^{\varepsilon_n}(i_n)|h_n) = 1, \forall h_n = (i_0, a_0, \dots, i_n) \in H_n, n \geq 0$, from (4.15) we can derive that

$$E_\pi^i[u_{n+1}(X_{n+1})] + E_\pi^i[r_n(X_n, A_n)] \geq g_n + E_\pi^i[u_n(X_n)] - \varepsilon_n. \quad (4.16)$$

Since $E_\pi^i[u_n(X_n)] < \infty, \forall n \geq 0$, by (4.16) we obtain

$$\bar{V}(\pi, i) \geq \rho^* - \varepsilon. \quad (4.17)$$

Since $\pi \in \Pi^*(\varepsilon_n)$, we can derive

$$0 \leq -D_n(X_n, A_n) \leq \varepsilon_n, n \geq 0, \quad a.e. -P_\pi^i. \quad (4.18)$$

From (4.12), we have, a.e. $-P_\pi^i$

$$\begin{aligned} 0 &= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} [r_n(X_n, A_n) - g_n - D_n(X_n, A_n)] \\ &\leq \liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} \left[r_n(X_n, A_n) - \frac{1}{N} \sum_{n=0}^{N-1} g_n + \varepsilon_n \right] \\ &\leq \limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} [r_n(X_n, A_n)] + \liminf_{N \rightarrow \infty} \left[-\frac{1}{N} \sum_{n=0}^{N-1} g_n \right] + \limsup_{N \rightarrow \infty} \frac{\sum_{n=0}^{N-1} \varepsilon_n}{N} \\ &= \limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} [r_n(X_n, A_n)] - \limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} g_n + \varepsilon. \end{aligned}$$

Hence,

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} [r_n(X_n, A_n)] \geq \limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} g_n - \varepsilon, \quad a.e. -P_\pi^i. \quad (4.19)$$

$$P_\pi^i \left\{ \limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} [r_n(X_n, A_n)] \geq \limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} g_n - \varepsilon \right\} = 1, \quad \forall i \in S_0. \quad (4.20)$$

Taking expectation operator E_π^i on the two sides of (4.19), one has

$$\begin{aligned} \bar{U}(\pi, i) &= E_\pi^i \limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} [r_n(X_n, A_n)] \\ &\geq \limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} g_n - \varepsilon. \end{aligned} \quad (4.21)$$

From Theorem 4.1, together with (4.17), (4.20) and (4.21), we have that part (i) is valid. We now prove part (ii). Since $D_n(X_n, \Delta_n) = 0$, *a.e.* $-P_{\pi^*}^i$ for all $i \in S_0$, from (4.12), we have

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} [r_n(X_n, \Delta_n) - g_n] = 0. \tag{4.22}$$

Hence, for all $i \in S_0$

$$\begin{aligned} & \left| \limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} [r_n(X_n, \Delta_n)] - \limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} g_n \right| \\ & \leq \limsup_{N \rightarrow \infty} \left| \frac{1}{N} \sum_{n=0}^{N-1} [r_n(X_n, \Delta_n) - g_n] \right| \\ & = \lim_{N \rightarrow \infty} \left| \frac{1}{N} \sum_{n=0}^{N-1} [r_n(X_n, \Delta_n) - g_n] \right| = 0 \quad \textit{a.e.} \quad -P_{\pi^*}^i, \end{aligned} \tag{4.23}$$

therefore,

$$\limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} [r_n(X_n, \Delta_n)] = \limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} g_n, \quad \textit{a.e.} \quad -P_{\pi^*}^i. \tag{4.24}$$

$$P_{\pi^*}^i \left\{ V_s(\pi^*, i) = \limsup_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} g_n \right\} = 1. \tag{4.25}$$

By Theorem 4.1, (4.24) and (4.25), we can obtain part (a) is valid. By Theorem 4.1 and (a), we conclude that the policy is strong average optimal. Hence, to complete the proof of part (ii), we only need to show that the policy is canonical. For the policy π^* , we have $Z_n(X_n, \Delta_n) = 0$, *a.e.* $-P_{\pi^*}^i$ for all $i \in S_0$. Hence, as the proof of (4.8), we can derive that

$$E_{\pi^*}^i [u_{N+1}(X_{N+1})] - u_0(i) + \sum_{n=0}^N E_{\pi^*}^i [r_n(X_n, \Delta_n)] = \sum_{n=0}^N g_n. \tag{4.26}$$

From Theorem 4.1 and (4.26), we have that (g_n, u_n, π^*) is a canonical triplet, therefor the policy π^* is indeed canonical. ▽▽▽

To ensure the existence of strong average-canonical optimal Markov policies and analyze the properties of optimal policies, we provide the following conditions

Assumption 4.1. $\sum_{n=1}^{\infty} \frac{\|R_n\|^2}{n^2} < \infty$.

Also, we need the following definition:

Definition 4.1. A policy $\pi \in \Pi$ is said to be a convex combination of two policies π^1 and π^2 in Π , if there exists a sequence of number $p_n \in [0, 1]$ such that

$$\pi(\cdot|h_n) = p_n\pi^1(\cdot|h_n) + (1 - p_n)\pi^2(\cdot|h_n), \quad \forall h_n \in H_n \quad \text{and} \quad n \geq 0.$$

Now we provide the main results in this paper.

Theorem 4.4. (i) If Assumptions (3.1), (3.2) and (4.1) are satisfied, then

- (a) there exists a strong average-canonical optimal Markov policy;
- (b) if $\pi^1, \pi^2 \in \Pi^*(\varepsilon)$, any convex combination of π^1 and π^2 is strong average ε optimal.
- (c) if $\pi^1, \pi^2 \in \Pi^*(0)$, any convex combination of π^1 and π^2 is strong average-canonical optimal.
- (ii) If every $A_n(i)$ ($i \in S_n, n \geq 0$) is finite, $\beta := \sup_n |(1 - \delta_n(S))| < 1$ and $\sup_n \|r_n\| < \infty$, then there exists a strong average-canonical optimal Markov policy.

Proof. By Theorems 3.1, 3.2, 4.2, the (a) of part (i) can be proved. the (b) and (c) in part (i) follow from Theorem 4.3. The (a) of part (i) yields part (ii).

▽▽▽

We now discuss some properties about the average expected criterion \bar{V} studied in [18, 1, 3, 20, 19, 11, 12, 13].

In order to do so, we let $\{g_n\}$ and $\{u_n\}$ be a solution to the OE's (3.1), for any $h = (i_0, a_0, \dots, i_n, a_n, \dots) \in H_\infty, i_n \in S_n, a \in A_n(i)$ and $n \geq 0$, define

$$M_0(i_0) := u_0(i),$$

$$M_n(h) := \sum_{t=0}^{n-1} r_t(i_t, a_t) + u_t(i_t) - \sum_{t=0}^{n-1} g_t \quad \forall n \geq 1. \tag{4.27}$$

Theorem 4.5. Suppose Assumption 3.1 holds, and

$$\lim_{N \rightarrow \infty} \frac{u_N(X_N)}{N} = 0, \tag{4.28}$$

for a policy $\pi \in \Pi$ and $i \in S_0$, then

- (i) $\lim_{N \rightarrow \infty} \frac{J_N(\pi, i) - \sum_{n=0}^{N-1} g_n}{N} = 0$ if and only if
- $$\lim_{N \rightarrow \infty} \frac{\sum_{n=0}^{N-1} E_\pi^i D_n(X_n, A_n)}{N} = 0; \tag{4.29}$$

- (ii) if (4.29) holds, then $\bar{V}(\pi, i) = \rho^*$;
- (iii) $\{M_n\}$ is a P_π^i -supermartingale;
- (iv) if $\{M_n\}$ is a P_π^i -martingale, then $\bar{V}(\pi, i) = \rho^*$.

Proof. From (2.3) and (3.1), we obtain, for all $n \geq 0$

$$E_{\pi}^i[D_n(X_n, \mathcal{A}_n)|X_0, \mathcal{A}_0, \dots, X_n, \mathcal{A}_n] \tag{4.30}$$

$$= r_n(X_n, \mathcal{A}_n) + \sum_{j \in S_{n+1}} u_{n+1}(j)P_n(j|X_n, \mathcal{A}_n) - u_n(X_n) - g_n, \tag{4.31}$$

which gives

$$E_{\pi}^i[D_n(X_n, \mathcal{A}_n)] = E_{\pi}^i[r_n(X_n, \mathcal{A}_n) + u_{n+1}(X_{n+1}) - u_n(X_n) - g_n].$$

Therefore,

$$\sum_{n=0}^{N-1} E_{\pi}^i[D_n(X_n, \mathcal{A}_n)] = J_N(\pi, i) + E_{\pi}^i u_N(X_N) - u_0(i) - \sum_{n=0}^{N-1} g_n. \tag{4.32}$$

Multiplying (4.32) by $1/N$ and let $N \rightarrow \infty$, we obtain part (i). From (4.32), we have

$$\sum_{n=0}^{N-1} E_{\pi}^i[D_n(X_n, \mathcal{A}_n)] + \sum_{n=0}^{N-1} g_n = J_N(\pi, i) + E_{\pi}^i u_N(X_N) - u_0(i), \tag{4.33}$$

which yields part (ii).

(iii) For any n -history $h_n = (i_0, a_0, \dots, i_n)$, from (4.27) and (4.9), we can obtain that

$$E_{\pi}^i(M_{n+1}|h_n) = M_n + E_{\pi}^i(D_n(X_n, \mathcal{A}_n)|h_n), \quad \forall n \geq 0. \tag{4.34}$$

Thus, since D_n is nonpositive, we obtain part (iii).

Finally, if $\{M_n\}$ is a P_{π}^i -martingale, then $E_{\pi}^i(M_n) = E_{\pi}^i(M_0) = h(i)$, that is, from (4.27),

$$J_N(\pi, i) + E_{\pi}^i(u_N) = h(i) + \sum_{n=0}^{N-1} g_n,$$

which implies part (iv). ▽▽▽

Now we deal with the problem of optimal average value convergence. Let $J_N^*(i) := \sup_{\pi \in \Pi} J_N(\pi, i)$ be the optimal N -horizon reward value, $\bar{V}^*(i) := \sup_{\pi \in \Pi}$ be the infinite horizon optimal average reward value.

Theorem 4.6. *If Assumptions (3.1) holds and $\lim_{N \rightarrow \infty} \frac{R_N}{N} = 0$, then*

$$\limsup_{N \rightarrow \infty} \frac{J_N^*(i)}{N} = \bar{V}^*(i) = \rho^*, \quad \forall i \in S_0.$$

Proof. Under the Assumptions (3.1), let $\{g_n\}$ and $\{u_n\}$ be a solution to the OE's (3.1). Form (4.8), for any $\pi \in \Pi$ and $i \in S_0$, we have

$$J_N(\pi, i) \leq \sum_{n=0}^{N-1} g_n + u_0(i) - E_{\pi}^i[u_N(X_N)],$$

which gives

$$J_N^*(i) \leq \sum_{n=0}^{N-1} g_n + u_0(i) + R_N. \tag{4.35}$$

By (4.28) and (4.35), we have

$$\limsup_{N \rightarrow \infty} \frac{J_N^*(i)}{N} \leq \rho^* \quad \text{and} \quad \bar{V}^*(i) \leq \rho^*. \tag{4.36}$$

On the other hand, by Theorem 4.2 we have, for any $\varepsilon > 0$, there exists a Markov policy $\pi \in \Pi^*(\varepsilon)$. As the proof of (4.8), we have

$$J_N(\pi(\varepsilon), i) \geq \sum_{n=0}^{N-1} g_n + u_0(i) - E_{\pi(\varepsilon)}^i[u_N(X_N)] - N\varepsilon. \tag{4.37}$$

Hence,

$$J_N^*(i) \geq J_N(\pi(\varepsilon), i) \geq \sum_{n=0}^{N-1} g_n + u_0(i) - R_N - N\varepsilon, \tag{4.38}$$

which yields

$$\limsup_{N \rightarrow \infty} \frac{J_N(i)}{N} \geq \rho^* - \varepsilon \quad \text{and} \quad \bar{V}^*(i) \geq \rho^* - \varepsilon. \tag{4.39}$$

Let $\varepsilon \rightarrow 0$, from (4.39) and (4.36), we can finish the proof of this theorem.

▽▽▽

Remark 4.1. In [20], the result of this theorem has also been proved. But in [20] the following additional assumptions are needed: (1) from each state i , at stage n , under action $a \in A_n(i)$, there exists a finite set $\{j | P_n(j|i, a) > 0\}$. That is, only a finite set of states is reachable in one step transition from any state, under any action; (2) also, for every $n \geq 0$ and $i \in S_n \equiv S$, $A_n(i)$ is finite; (3) furthermore, the reward functions r_n are required to be uniformly bounded in n , i.e., $\sup_{n \geq 0} \|r_n\| < \infty$.

Next we will show that Assumptions 2 and 3 in [20] fail to hold in the following example, whereas, all assumptions introduced in this paper hold.

Example 4.1. Let $S_n \equiv S := \{1, 2, 3, 4, 5\}$ and $A_n(i)$ ($n \geq 0, i \in S$) be nonempty finite set. The rewards $r_n(i, a) := n^{1/4} \bar{r}_n(i, a)$, $\forall a \in A(i), i \in S, n \geq 0$, here, $\bar{r}_n(i, a) \in (m, M)$, $a \in A(i), i \in S, n \geq 0, 0 < m < M$. The transition law P_n is defined by

$$P_n := \begin{pmatrix} 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 \end{pmatrix}, \quad n \in N_1, \quad (4.40)$$

$$P_n := \begin{pmatrix} \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & 0 & \frac{2}{5} \\ \frac{1}{4} & \frac{1}{4} & 0 & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{6} & 0 & \frac{1}{3} & \frac{1}{6} & \frac{1}{3} \\ \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} & \frac{1}{5} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \end{pmatrix}, \quad n \in N_2, \quad (4.41)$$

where, $N_1 \cup N_2 = \{0, 1, \dots, n, \dots\}$, $N_1 \cap N_2 = \emptyset$, and N_1 or N_2 may be empty. Obviously, the rewards r_n is unbounded in time n , i.e. $\sup_{n \geq 0} \|r_n\| = \infty$. Hence, for this example we have that: (1) the assumption that $\sup_{n \geq 0} \|r_n\| < \infty$ in [18, 1, 3, 20, 19, 11, 12, 13] fails to hold, and (2) if $N_1 \neq \emptyset$, then $\beta := \sup_n (1 - \sum_{j \in S} \inf_{a \in A_n(i)} P_n(j|i, a)) = 1$, therefor, the condition that $\beta < 1$ in [20, 13] is not satisfied. But, we will show that all Assumptions in this paper do hold.

In fact, we take $\delta_n(j) := \max_{i \in S a \in A(i)} P_n(j|i, a)$, $\forall j \in S, n \in N_1$, and, for all $j \in S, n \in N_2$, let $\delta_n(j) := \min_{i \in S a \in A(i)} P_n(j|i, a)$. Then we have $\delta_n(S) = \frac{5}{4}$, $\forall n \in N_1$, and $\delta_n(S) = \frac{1}{6}$, $\forall n \in N_2$. Hence, $\beta^* := \sup_{n \geq 0} (1 - \delta_n(S)) = \frac{5}{6} < 1$.

From $\beta^* < 1$, we can easily verify that

$$\begin{aligned} \|R_n\| &\leq n^{1/4} M + \sum_{k=1}^{\infty} \beta^{*k} (n+k)^{1/4} M \\ &\leq n^{1/4} M + \sum_{k=1}^{\infty} \beta^{*k} [n^{1/4} + k^{1/4}] M \\ &\leq n^{1/4} \left[1 + \frac{\beta^*}{1 - \beta^*} \right] M + \sum_{k=1}^{\infty} \beta^{*k} k^{1/4} M < \infty. \end{aligned} \quad (4.42)$$

Hence, Assumption 3.1 holds. From (4.42), we easily have that $\sum_{n=1}^{\infty} \frac{\|R_n\|^2}{n^2} < \infty$. This means that Assumption 4.1 holds. Hence, all Assumptions made in this paper are satisfied for this example.

5 An algorithm

In this section, the error bound in strong average reward between a rolling horizon policy and a strong average-canonical optimal policy is provided, and then a rolling horizon algorithm for computing strong average $\varepsilon(>0)$ -optimal Markov policies is proposed.

For a fixed finite horizon $N_0 \geq 1$, the Rolling Horizon Algorithm is stated as follows:

Step 1. Set $m = 0$ and $n = N_0$.

Step 2. Find a $f_k^*(N_0), m \leq k \leq m + N_0$ such that the so-called finite horizon Markov policy $\pi^* := (f_m^*(N_0), \dots, f_{m+N}^*(N_0)^*)$ is optimal for periods m through n , and set $\hat{f}_m(N_0) := f_m^*(N_0)$.

Step 3. Let $m = m + 1$ and $N = N_0 + 1$.

Step 4. Go to Step 2.

The algorithm above recursively generates the infinite horizon policy:

$$\hat{\pi}(N_0) := (\hat{f}_0(N_0), \hat{f}_1(N_0), \dots, \hat{f}_n(N_0), \dots).$$

The policy $\hat{\pi}(N_0)$ is called a rolling N_0 -horizon algorithm.

Under Assumption 3.1, for $n \geq 0, i \in S_n$, let

$$\begin{aligned} V_{n+N_0+1}(\cdot) &:= \max_{a \in A_{n+N_0+1}(\cdot)} r_{n+N_0+1}(\cdot, a), \\ V_{n+N_0}(\cdot) &:= \max_{a \in A_{n+N_0}(\cdot)} \left[r_{n+N_0}(\cdot, a) + \sum_{j \in S_{N_0+n}} (P_{n+N_0}(j|\cdot, a) - \delta_{n+N_0}(j)) V_{n+N_0+1}(j) \right] \\ &\vdots \\ &\vdots \\ V_n(i) &:= \max_{a \in A_n(i)} \left[r_n(i, a) + \sum_{j \in S_{n+1}} (P_n(j|i, a) - \delta_n(j)) V_{n+1}(j) \right]. \end{aligned} \quad (5.1)$$

From (3.6), (3.8) and (3.10), we know that

$$\begin{aligned} |V_n(i) - u_n(i)| &\leq \sum_{k=N_0+1}^{\infty} |(1 - \delta_n(S_{n+1})) \cdots (1 - \delta_{n+k-1}(S_{n+k}))| \|r_{n+k}\| \\ &:= \bar{R}_n(N_0). \end{aligned} \quad (5.2)$$

For any $i \in S_n$ and $n \geq 0$, we may choose $f_n^{N_0}(i) \in A_n(i)$ such that

$$V_n(i) = r_n(i, f_n^{N_0}(i)) + \sum_{j \in S_{n+1}} (P_n(j|i, f_n^{N_0}(i)) - \delta_n(j)) V_{n+1}(j). \quad (5.3)$$

From (5.2) and (5.3), one has

$$\begin{aligned} &r_n(i, f_n^{N_0}(i)) + \sum_{j \in S_{n+1}} P_n(j|i, f_n^{N_0}(i)) u_{n+1}(j) \\ &\geq u_n(i) + g_n - \bar{R}_n(N_0) - |1 - \delta_n(S)| \bar{R}_{n+1}(N_0) \\ &:= u_n(i) + g_n - \varepsilon_n(N_0), \end{aligned} \quad (5.4)$$

where $\varepsilon_n(N_0) := \bar{R}_n(N_0) - |1 - \delta_n(S)| \bar{R}_{n+1}(N_0)$.

Let $\hat{\pi}(N_0) := \{f_n^{N_0}\}$. From (5.4), under Assumptions 3.1 and 4.1, by Theorem 4.1 we have

$$\bar{V}(\hat{\pi}(N_0), i) \geq \rho^* - \varepsilon(N_0); \tag{5.5}$$

$$\bar{U}(\hat{\pi}(N_0), i) \geq \rho^* - \varepsilon(N_0); \tag{5.6}$$

$$P_{\hat{\pi}(N_0)}^i \{V_s(\hat{\pi}(N_0), i) \geq \rho^* - \varepsilon(N_0)\} = 1, \tag{5.7}$$

here, $\varepsilon(N_0) := \limsup_{N \rightarrow \infty} \frac{\sum_{n=0}^{N-1} \varepsilon_n(N_0)}{N}$.

From the above discussions, we have the following theorem:

Theorem 5.1. *If Assumptions 3.1 and 4.1 are satisfied, then the errors in average expected reward between a rolling N_0 -horizon policy $\hat{\pi}(N_0)$ and a strong average-canonical optimal policy π^* satisfy (5.8), (5.9) and (5.10).*

$$\bar{V}(\pi^*, i) - \bar{V}(\hat{\pi}(N_0), i) \leq \varepsilon(N_0); \tag{5.8}$$

$$\bar{U}(\pi^*, i) - \bar{U}(\hat{\pi}(N_0), i) \leq \varepsilon(N_0); \tag{5.9}$$

$$P_{\hat{\pi}(N_0)}^i \{V_s(\pi^*, i) - V_s(\hat{\pi}(N_0), i) \leq \varepsilon(N_0)\} = 1. \tag{5.10}$$

From Theorem 5.1, the following corollary follows.

Corollary 5.1. *If the following conditions hold*

(i) *Assumption 3.1 holds, $\beta^* := \sup_{n \geq 0} |(1 - \delta_n(S_{n+1}))| < 1$, and $M := \sup_{n \geq 0} \|r_n\| < \infty$;*

(ii) *for any fixed $n \geq 0, i \in S_n$ and $N_0 \geq 1$, the maximum point of (5.1) can be obtained in finite steps.*

Then, for any fixed $\varepsilon > 0$, and given $n \geq 0, i \in S_n$, the action $f_n^(i)$ can be calculated in finite steps and the corresponding Markov policy $\pi^* = \{f_n^*\}$ is strong average $\varepsilon(N_0)$ -optimal.*

Proof. Under the condition (i), for any $\varepsilon > 0$, we can choose a positive integer N_0 such that $\beta^{*N_0} \frac{M}{1 - \beta^*} < \frac{\varepsilon}{4}$. By (5.2), we have $\varepsilon_n(N_0) \leq \varepsilon, \forall n \geq 0$. Hence, by condition (ii), for any fixed $n \geq 0$ and $i \in S$, there exists $f_n(i) \in A_n(i)$ such that

$$r_n(i, f_n^*(i)) + \sum_{j \in S} P_n(j|i, f_n^*(i)) u_{n+1}(j) \geq u_n(i) + g_n - \varepsilon. \tag{5.11}$$

Hence, by Theorem 4.3, we have that the policy $\hat{\pi}(N_0) := \{f_n^*\}$ is strong average ε -optimal.

On the other hand, since $\sum_{j \in S} V_k(j) \delta_{k+1}(j)$ is independent of both action a and state i for any $k \geq 0$, by induction, we can prove that the $f_n^{N_0}(i)$ satisfying (5.3) may be obtained from the following (5.14):

$$W_{n+N_0+1}(\cdot) := \max_{a \in A_{n+N_0+1}(\cdot)} r_{n+N_0+1}(\cdot, a), \tag{5.12}$$

$$W_{n+N_0}(\cdot) := \max_{a \in A_{n+N_0}(\cdot)} \left[r_{n+N_0}(\cdot, a) + \sum_{j \in S_{n+N_0+1}} P_{n+N_0}(j|\cdot, a) W_{n+N_0+1}(j) \right]$$

$$\vdots$$

$$W_n(i) := \max_{a \in A_n(i)} \left[r_n(i, a) + \sum_{j \in S_{n+1}} P_n(j|i, a) W_{n+1}(j) \right]. \tag{5.13}$$

$$W_n(i) = r_n(i, f_n^{N_0}(i)) + \sum_{j \in S_{n+1}} P_n(j|i, f_n^{N_0}(i)) W_{n+1}(j). \tag{5.14}$$

Finally, we can provide an algorithm to find an strong average $\varepsilon > 0$ -optimal Markov policy $\pi^* = \{f_n^{N_0}\}$ as follows:

- Step 1** for $\varepsilon > 0$, choose a positive integer N_0 such that $\beta^{N_0} \frac{M}{1 - \beta^*} < \frac{\varepsilon}{4}$;
- Step 2** for given $n \geq 0$ and $i \in S_n$, by (5.13), calculate $W_n(i)$;
- Step 3** select $f_n^*(i) \in A_n(i)$ satisfying (5.14).

Remark 5.1. *The above algorithm is similar to the one given by Alden and Smith in [1]. However, the algorithm in [1] is restricted to the case of finite state space. In addition, the proof of convergence of our algorithm is different from that in [1].*

Acknowledgment. The authors wish to thank Professor Jerzy Filar for a number of discussions and his helpful suggestions. The first author is grateful to the hospitality during his visit in Centre for industrial and Applied Mathematics, The University of South Australia.

References

- [1] Alden M, Smith RL (1992) Rolling horizon procedures in nonhomogeneous Markov decision processes. *Oper. Res.* 40:183–194
- [2] Arapostathis A, Borcker VS, Fernandez-Gaucherand E, Ghosh MK, Marcus SI (1993) Discrete-time controlled Markov process with an average cost criterion: A survey. *SIAM J. on Control and Optim.* 31:282–344
- [3] Bean J, Smith R, Lasserre J (1990) Denumerable state nonhomogeneous Markov decision processes. *J. Math. Anal. Appl.* 153:64–77
- [4] Bierth KJ (1987) An expected average reward criterion. *Stoch. Pro. Appl.* 26:123–140
- [5] Blackwell D (1962) Discrete dynamic programming. *Ann. Math. Stat.* 33:719–726
- [6] Derman C (1970) Finite state Markov decision processes. Academic Press, New York
- [7] Dynkin EB, Yushkevich AA (1979) Controlled Markov processes. Springer-Verlag, New York
- [8] Feinberg E, Park H (1994) Finite state Markov decision model with average reward criteria. *Stoch. Pro. Appl.* 49:159–177
- [9] Filar JA, Krass D, Ross KW (1995) Percentile performance criteria for limiting average Markov decision processes. *IEEE Trans. Autom. Control* 40(1):2–10
- [10] Filar JA, Vrieze K (1997) Competitive Markov decision processes. Springer-Verlag New York, INC

- [11] Guo XP (1995) Nonstationary MDP average model with incomplete information. *J. Math. Statist. Appl. Prob.* 10(2):15–23
- [12] Guo XP (1998) The unique properties of optimal policies in general Markov decision processes. *J. Appl. Prob. and Stat.* 12(2):258–265
- [13] Guo XP (1999) Denumerable state nonhomogeneous Markov decision processes with average variance criterion. *ZOR-Math. Meth. Oper. Res.* 49(2):87–96
- [14] Guo XP, Liu JY, Liu K (2001) Nonstationary Markov decision processes with Borel state space – the average criterion with non-uniformly bounded rewards. *Math. Oper. Res.* to appear
- [15] Hernandez-Lerma O (1989) Adaptive Markov controlled processes. Springer-Verlag, New York
- [16] Hernandez-Lerma O, Lasserre JB (1996) Discrete-time Markov controlled processes. Springer-Verlag, New York
- [17] Hinderer K (1970) Foundations of non-stationary dynamic programming with discrete time parameter. Springer-Verlag, New York
- [18] Hopp W, Bean J, Smith R (1987) A new optimality criterion for nonhomogeneous Markov decision processes. *Oper. Res.* 35:875–883
- [19] Hou ZT, Guo XP (1998) Markov decision processes. Science and Technology Press, Hunan, China
- [20] Park Y, Bean J, Smith R (1993) Optimal average value convergence in nonhomogeneous Markov decision processes. *J. Math. Anal. Appl.* 179:525–536
- [21] Puterman M (1994) Markov decision processes: Discrete stochastic dynamic programming. John Wiley & Sons Inc
- [22] Roland CC, Emanuel FG (1995) Denumerable controlled Markov chains with average reward criterion: sample path optimality. *Z. Oper. Res.* 41(2):89–108
- [23] Ross KW, Varadarajan R (1989) Markov decision processes with sample path constraints: the communicating case. *Oper. Res.* 37:780–790
- [24] Ross KW, Varadarajan R (1991) Markov decision processes with sample path constraints: A decomposition approach. *Math. Ope. Res.* 16(1):195–207
- [25] Shiriyayev AN (1984) Probability. Springer-Verlag, New York Berlin Heidelberg Tokyo