A LEARNING ALGORITHM FOR COMMUNICATING MARKOV DECISION PROCESSES WITH UNKNOWN TRANSITION MATRICES

by

Tetsuichiro Iki, **Masayuki Horiguchi**, **Masami Yasuda**
and
Masami Kurano

Reprinted from the Bulletin of Informatics and Cybernetics
Research Association of Statistical Sciences, Vol.39

FUKUOKA, JAPAN
2007

Bulletin of Informatics and Cybernetics, Vol. 39, 2007

A LEARNING ALGORITHM FOR COMMUNICATING MARKOV DECISION PROCESSES WITH UNKNOWN TRANSITION MATRICES

By

Tetsuichiro Iki[*], **Masayuki Horiguchi**[†], **Masami Yasuda**[‡]
and
Masami Kurano[§]

Abstract

This study is concerned with finite Markov decision processes (MDPs) whose state are exactly observable but its transition matrix is unknown. We develop a learning algorithm of the reward-penalty type for the communicating case of multi-chain MDPs. An adaptively optimal policy and an asymptotic sequence of adaptive policies with nearly optimal properties are constructed under the average expected reward criterion. Also, a numerical experiment is given to show the practical effectiveness of the algorithm.

Key Words and Phrases: Adaptive policy, Average case, Communicating case, Learning algorithm, Markov decision processes, Reward-penalty type, Unknown transition matrix.

1. Introduction and notation

In the real world, there are many requests to solve uncertain models. Adaptive models for uncertain Markov decision processes (MDPs) have been considered by many authors as Hernández (1989), Hernández/ Marcus (1985), Kurano (1972), Kurano (1983), Mandl (1974), Martin (1967), van Hee (1978) and so on. The idea of Neuro dynamic programming by Bertsekas/Tsitsiklis (1996) is powerful in treating the adaptive MDPs.

However, a simple learning algorithm of the reward-penalty type, investigated by Lakshmivarahan (1981) and Meybodi/Lakshmivarahan (1982) are more comprehensible and manageable. Kurano (1987) proposed a learning algorithms of the reward-penalty type where all elements of the true transition matrices of finite MDPs are known to be positive and constructed adaptively optimal policy under the average expected reward criterion.

In this paper, applying the idea of Kurano (1987) to a wider class of uncertain MDPs, we develop a learning algorithm for the communicating case of multi-chain MDPs

[*] Faculty of Education and Culture, Miyazaki University, Miyazaki 889-2192 Japan, e03101u@cc.miyazaki-u.ac.jp
[†] General Education, Yuge National College of Maritime Technology, Ehime 794-2593 Japan, horiguchi@gen.yuge.ac.jp
[‡] Faculty of Science, Chiba University, Chiba 263-8522 Japan, yasuda@math.s.chiba-u.ac.jp
[§] Faculty of Education, Chiba University, Chiba 263-8522 Japan, kurano@faculty.chiba-u.jp

and construct an adaptively average optimal policy for a class of perturbed communicating MDPs. For general communicating MDPs, an asymptotic sequence of adaptive policies with nearly optimal properties is constructed by using the results of perturbed case.

In the reminder of this section, we will formulate finite MDPs whose transition matrices are unknown but the state at each stage is observable exactly. Consider a controlled dynamic system with finite state and action spaces, S and A, containing $N < \infty$ and $K < \infty$ elements respectively. Let $\mathbb{Q}$ denote the parameter space of K unknown stochastic matrices, that is

$$\mathbb{Q} = \Big\{ q = (q_{ij}(a)) \;\Big|\; q_{ij}(a) \geqq 0, \sum_{j \in S} q_{ij}(a) = 1 \text{ for } i, j \in S, a \in A \Big\}.$$

The sample space is the product space $\Omega = (S \times A)^\infty$ such that the projections X_t, Δ_t on the t-th factors S, A describe the state and action at the t-th stage of the process($t \geqq 0$). Let Π denote the set of all policies, i.e., for $\pi = (\pi_0, \pi_1, \ldots) \in \Pi$, let $\pi_t \in P(A|(S \times A)^t \times S)$ for all $t \geqq 0$, where, for any finite sets X and Y, $P(X|Y)$ denotes the set of all conditional probability distribution on X given Y. A policy $\pi = (\pi_0, \pi_1, \ldots)$ is called randomized stationary if a conditional probability $\gamma = (\gamma(\cdot|i) : i \in S) \in P(A|S)$ such that $\pi_t(\cdot|x_0, a_0, \ldots, x_t) = \gamma(\cdot|x_t)$ for all $t \geqq 0$ and $(x_0, a_0, \ldots, x_t) \in (S \times A)^t \times S$. Such a policy is simply denoted by γ. We denote by F the set of functions on S with $f(i) \in A$ for all $i \in S$. A randomized stationary policy γ is called stationary if there exists a function $f \in F$ with $\gamma(\{f(i)\}|i) = 1$ for all $i \in S$, which is denoted simply by f.

We will construct a probability space as follows: For any initial state $X_0 = i, \pi \in \Pi$ and a transition law $q = (q_{ij}(a)) \in \mathbb{Q}$, let $P(X_{t+1} = j|X_0, \Delta_0, \ldots, X_t = i, \Delta_t = a) = q_{ij}(a)$ and $P(\Delta_t = a|X_0, \Delta_0, \ldots, X_t = i) = \pi_t(a|X_0, \Delta_0, \ldots, X_t = i)$ ($t \geqq 0$). Then, we can define the probability measure $P_\pi(\cdot|X_0 = i, q)$ on Ω. For a given reward function r on $S \times A$, we shall consider the long-run expected average reward:

$$\psi(i, q|\pi) = \liminf_{T \to \infty} \frac{1}{T+1} E_\pi \Big(\sum_{t=0}^{T} r(X_t, \Delta_t) \;\Big|\; X_0 = i, q \Big) \tag{1.1}$$

where $E_\pi(\cdot|X_0 = i, q)$ is the expectation operator with respect to $P_\pi(\cdot|X_0 = i, q)$.

Let $\mathcal{D}$ be a subset of $\mathbb{Q}$. Then, the problem is to maximize $\psi(i, q|\pi)$ over all $\pi \in \Pi$ for any $i \in S$ and $q \in \mathcal{D}$. Thus, denoting the optimal value function as

$$\psi(i, q) = \sup_{\pi \in \Pi} \psi(i, q|\pi), \tag{1.2}$$

a policy $\pi^* \in \Pi$ will be called q-optimal if $\psi(i, q|\pi^*) = \psi(i, q)$ for all $i \in S$ and called adaptively optimal for $\mathcal{D}$ if π^* is q-optimal for all $q \in \mathcal{D}$. A sequences of policies $\{\pi^n\}_{n=1}^\infty \subset \Pi$ is called an asymptotic sequence of adaptive policies with nearly optimal properties for $\mathcal{D}$ if

$$\lim_{n \to \infty} \psi(i, q|\pi^n) = \psi(i, q) \quad \text{for all } q \in \mathcal{D}.$$

In Kurano (1987), an adaptively optimal policy for $\mathbb{Q}^+ := \{q = (q_{ij}(a)) \in \mathbb{Q} \;|\; q_{ij}(a) > 0 \text{ for all } i, j \in S \text{ and } a \in A\}$ was constructed by applying the value iteration and the policy improvement algorithm (cf. Lakshmivarahan (1981), Federgruen/Schweitzer (1981),

Hernández (1989), Hernández/ Marcus (1985)). In this paper, we treat with the communicating case of multi-chain MDPs applying the idea of Kurano (1987) extensively.

A transition matrix $q = (q_{ij}(a)) \in \mathbb{Q}$ is said communicating (cf. Bather (1973), Puterman (1994)) if for any $i, j \in S$ there exists a path from i to j with positive probability, i.e., it holds that

$$q_{i_1 i_2}(a_1) q_{i_2 i_3}(a_2) \cdots q_{i_{l-1} i_l}(a_{l-1}) > 0$$

for some $\{i_1 = i, i_2, \ldots, i_l = j\} \subset S$ and $\{a_1, a_2, \ldots, a_{l-1}\} \subset A$ and $2 \leqq l \leqq N$. It is easily shown that $q = (q_{ij}(a))$ is communicating if and only if there is a randomized stationary policy $\gamma = (\gamma(\cdot|i) : i \in S)$ satisfying that the transition matrix $q(\gamma) = (q_{ij}(\gamma))$ induced by γ defines an irreducible Markov chain (cf. Kemeny/Snell (1960)) where $q_{ij}(\gamma) = \sum_{a \in A} q_{ij}(a)\gamma(a|i)$ for $i, j \in S$.

Let $B(S)$ be the set of all functions on S. The following fact is well-known (cf. Puterman (1994), Ross (1970)).

Lemma 1.1 (Puterman (1994), Ross (1970)) *Let $q = (q_{ij}(a)) \in \mathbb{Q}$. Supposed that there exists a constant g and a $v \in B(S)$ such that, for all $i \in S$,*

$$v(i) = \max_{a \in A} \Big\{ r(i,a) + \sum_{j \in S} q_{ij}(a) v(j) \Big\} - g. \tag{1.3}$$

Then, g is unique and $g = \psi(i,q) = \psi(i,q|f)$ for $i \in S$, where $f \in F$ is q-optimal and $f(i)$ is a maximizer in the right-hand side of (1.3) for all $i \in S$.

Let $\mathbb{Q}^*$ be the set of all communicating transition matrices. In order to treat with the communicating case with $q \in \mathbb{Q}^*$, we use the so-called vanishing discount approach which studies the average case by considering the corresponding $(1-\tau)$-discounted one as letting $\tau \to 0$. The expected total $(1-\tau)$-discounted reward is defined by

$$v_\tau(i, q|\pi) = E_\pi \Big(\sum_{t=0}^{\infty} (1-\tau)^t r(X_t, \Delta_t) \;\Big|\; X_0 = i, q \Big) \tag{1.4}$$

for $i \in S, q \in \mathbb{Q}$ and $\pi \in \Pi$, and $v_\tau(i,q) = \sup_{\pi \in \Pi} v_\tau(i, q|\pi)$ is called a $(1-\tau)$-discounted value function, where $(1-\tau) \in (0,1)$ is a given discount factor.

For any $q = (q_{ij}(a)) \in \mathbb{Q}$ and $\tau \in (0,1)$, we define the operator $U_\tau\{q\} : B(S) \to B(S)$ by

$$U_\tau\{q\}u(i) = \max_{a \in A} \Big\{ r(i,a) + (1-\tau) \sum_{j \in S} q_{ij}(a) u(j) \Big\} \tag{1.5}$$

for all $i \in S$ and $u \in B(S)$. We have the following.

Lemma 1.2 (Puterman (1994), Ross (1970)) *It holds that*

(i) *the operator $U_\tau\{q\}$ is a contraction with the modulus $(1-\tau)$,*

(ii) *the $(1-\tau)$-discount value function $v_\tau(i,q)$ is a unique fixed point of $U_\tau\{q\}$, i.e.,*

$$v_\tau = U_\tau\{q\} v_\tau, \tag{1.6}$$

(iii) $v_\tau(i,q) = v_\tau(i,q|f_\tau)$ *and* $\lim_{\tau\to 0} \tau v_\tau(i,q) = \psi(i,q)$, *where* f_τ *is a maximizer of the right-hand side in* (1.6).

In Section 2, continuity of the value function for perturbed transition matrices is proved, by which an adaptively optimal policy for the perturbed communicating MDPs is constructed through a learning algorithm of reward-penalty type in Section 3. Also, Section 3 is devoted to the construction of an asymptotic sequence of adaptive policies with nearly optimal properties. In Section 4, a numerical experiment is implemented to show the practical effectiveness of the learning algorithm given in Section 3.

2. Continuity of the value function

First we give a key lemma for guaranteeing the validity of the vanishing discount approach to study the average case.

Lemma 2.1 *Let* $q = (q_{ij}(a)) \in \mathbb{Q}^*$. *Then, there exists a constant* M *such that*

$$\limsup_{\tau\to 0} \left| v_\tau(i,q) - v_\tau(j,q) \right| \leqq M \quad \textit{for all} \quad i,j \in S. \tag{2.1}$$

Proof. We denote by $H_t := (X_0, \Delta_0, \ldots, X_t)$ the history of states and actions until the t-th step($t \geqq 1$) with $H_0 = (X_0)$. For each $j \in S$, we define the stopping time σ^j by

$$\sigma^j = \sigma^j(H_t) = \text{first } t \geqq 0 \text{ such that } X_t = j.$$

That $q \in Q^*$ guarantees that there exists a randomized stationary policy $\gamma = (\gamma(\cdot|i) : i \in S)$ such that the Markov chain induced by $q(\gamma)$ is irreducible. Here, using the stationary policy f_τ given in Lemma 1.2 the policy $\pi^j = (\pi_0^j, \pi_1^j, \ldots)$ will be defined by

$$\pi_t^j(\cdot|H_t) = \begin{cases} \gamma(\cdot|X_t) & \text{if } t < \sigma^j(H_t), \\ f_\tau(X_t) & \text{if } t \geqq \sigma^j(H_t). \end{cases}$$

for $t \geqq 0$. Then we have the following: For $i \in S$,

$$v_\tau(i,q|\pi^j) = E_\gamma\Big(\sum_{t=0}^{\sigma^j - 1} (1-\tau)^t r(X_t, \Delta_t) \;\Big|\; X_0 = i, q \Big) + E_\gamma\Big((1-\tau)^{\sigma^j} \;\Big|\; X_0 = i, q \Big) v_\tau(j|q). \tag{2.2}$$

From irreducibility of the Markov chain induced by $q(\gamma)$, it holds (cf. Kemeny/Snell (1960)) that

$$E_\gamma\Big(\sigma^j \;\Big|\; X_0 = i, q \Big) < \infty \quad \text{for all} \quad i \in S. \tag{2.3}$$

Concerning with the second term of the right-hand side in (2.2), since $\lim_{\tau\to 0}\frac{(1-\tau)^n-1}{\tau} = -n$ $(n \geqq 1)$, we have that

$$\begin{aligned}
&\liminf_{\tau\to 0}\frac{1}{\tau}\Big\{E_\gamma\Big((1-\tau)^{\sigma^j}\ \Big|\ X_0=i,q\Big)-1\Big\}\\
&\geqq \sum_{n=0}^{\infty}\liminf_{\tau\to 0}\frac{(1-\tau)^n-1}{\tau}P_\gamma\Big(\sigma^j=n\ \Big|\ X_0=i,q\Big)\\
&= -\sum_{n=1}^{\infty} nP_\gamma\Big(\sigma^j=n\ \Big|\ X_0=i,q\Big) = -E_\gamma\Big(\sigma^j\ \Big|\ X_0=i,q\Big). \qquad (2.4)
\end{aligned}$$

On the other hand, from (2.2) it holds that

$$\begin{aligned}
v_\tau(i,q)-v_\tau(j,q) &\geqq v_\tau(i,q|\pi^j)-v_\tau(j,q)\\
&\geqq -\|r\|E_\gamma\Big(\sigma^j\ \Big|\ X_0=i,q\Big)+\Big\{E_\gamma\Big((1-\tau)^{\sigma^j}\ \Big|\ X_0=i,q\Big)-1\Big\}v_\tau(j,q)
\end{aligned}$$

where $\|r\| = \max_{i\in S,a\in A}|r(i,a)|$. Thus, by (2.2), (2.4) and Lemma 1.2(iii) we have that

$$\begin{aligned}
\liminf_{\tau\to 0}\Big(v_\tau(i,q)-v_\tau(j,q)\Big) &\geqq -\limsup_{\tau\to 0}\Big(\|r\|+|\tau v_\tau(j,q)|\Big)E_\gamma\Big(\sigma^j\ \Big|\ X_0=i,q\Big)\\
&= -\Big(\|r\|+|\psi(j,q)|\Big)E_\gamma\Big(\sigma^j|X_0=i,q\Big) > -\infty.
\end{aligned}$$

Similarly, we get that

$$\liminf_{\tau\to 0}\Big(v_\tau(j,q)-v_\tau(i,q)\Big) \geqq -\Big(\|r\|+|\psi(i,q)|\Big)E_\gamma\Big(\sigma^i|X_0=j,q\Big) > -\infty,$$

and hence

$$\limsup_{\tau\to 0}\Big|v_\tau(i,q)-v_\tau(j,q)\Big| \leqq \Big(\|r\|+|\psi(i,q)|\Big)E_\gamma\Big(\sigma^i|X_0=j,q\Big) < \infty.$$

If we put $M := \max_{i,j\in S}\Big(\|r\|+|\psi(j,q)|\Big)E_\gamma\Big(\sigma^j\ \Big|\ X_0=i,q\Big)$, (2.1) follows, which completes the proof. □

Let $P(S)$ be the set of all probability distributions on S, i.e.,

$$P(S)=\Big\{\mu=(\mu_1,\mu_2,\dots,\mu_N)\ \Big|\ \mu_i\geqq 0,\sum_{i=1}^{N}\mu_i=1 \text{ for all } i\in S\Big\}.$$

Let $q=(q_{ij}(a))\in\mathbb{Q}$. For any $\tau\in(0,1)$ and $\mu=(\mu_1,\mu_2,\dots,\mu_N)\in P(S)$, we perturb q to $q^{\tau,\mu}=(q_{ij}^{\tau,\mu}(a))$ which is defined by

$$q_{ij}^{\tau,\mu}(a)=\tau\mu_j+(1-\tau)q_{ij}(a) \quad \text{for } i,j\in S \text{ and } a\in A. \qquad (2.5)$$

The matrix expression of (2.5) is $q^{\tau,\mu}=\tau e\mu+(1-\tau)q$, where $e=(1,1,\dots,1)^t$ is a transpose of N-dimensional vector $(1,1,\dots,1)$. Then, we find that (1.6) in Lemma 1.2 can be rewritten as follows: For all $i\in S$,

$$v_\tau(i,q)=\max_{a\in A}\Big\{r(i,a)+\sum_{j\in S}q_{ij}^{\mu,\tau}(a)v_\tau(j,q)\Big\}-\tau\sum_{j\in S}\mu_j v_\tau(j,q). \qquad (2.6)$$

Thus, applying Lemma 1.1, we have the following.

Lemma 2.2 *For any $q \in \mathbb{Q}, \tau \in (0,1)$ and $\mu \in P(S)$, it holds that*

(i) $\psi(i, q^{\tau,\mu}) = \tau \sum_{j \in S} \mu_j v_\tau(j,q)$ *for all* $i \in S$,

(ii) f_τ *is* $q^{\tau,\mu}$*-optimal, where* f_τ *is given in Lemma 1.2.*

From Lemma 2.2, since $\psi(i, q^{\tau,\mu})$ is independent of $i \in S$, we shall put $\psi(q^{\tau,\mu}) := \psi(i, q^{\tau,\mu})$. The τ-continuity of $\psi(q^{\tau,\mu})$ is given in the following.

Theorem 2.1 *Let $q \in \mathbb{Q}^*$. Then, we have that*

(i) $\psi(i,q)(:= \psi(q))$ *is independent of $i \in S$ and there exists a $u \in B(S)$ satisfying the average optimality equation:*

$$u(i) = \max_{a \in A} \Big\{ r(i,a) + \sum_{j \in S} q_{ij}(a) u(j) \Big\} - \psi(q) \quad (i \in S), \tag{2.7}$$

(ii) *for any $\mu \in P(S), \psi(q^{\mu,\tau}) \to \psi(q)$ as $\tau \to 0$.*

Proof. For any fixed $i_0 \in S$, let $u_\tau(i) = v_\tau(i,q) - v_\tau(i_0,q)$ for each $i \in S$. Then, from (2.6) we get

$$u_\tau(i) = \max_{a \in A} \Big\{ r(i,a) + \sum_{j \in S} q_{ij}^{\mu,\tau}(a) u_\tau(j) \Big\} - \tau \sum_{j \in S} \mu_j v_\tau(j,q) \quad (i \in S). \tag{2.8}$$

By Lemma 1.2, $\lim_{\tau \to 0} \tau v_\tau(j,q) = \psi(j,q)$. Also, from Lemma 2.1, there exists a sequence $\{\tau_l\}$ with $\tau_l \to 0$ and $u_{\tau_l}(j) \to u(j)$ as $l \to \infty$ for some $u \in B(S)$ and all $j \in S$. Thus, letting $l \to \infty$ in (2.8) with $\tau = \tau_l$, we get (2.7) with $\psi(q) = \sum_{j \in S} \mu_j \psi(j,q)$. Applying Lemma 1.1, we observe that $\psi(q)$ is independent of $\mu \in P(S)$, so that (i) and (ii) follows. □

We note that (i) in Theorem 2.1 derives the single average optimality equation for the communicating MDPs, which has been given first by Bather (1973). In general, the value function $\psi(i,q)$ is known to be continuous on each equivalent class of $\mathbb{Q}$ (cf. Schweitzer (1968), Solan (2003)), but (ii) in Theorem 2.1 gives an example in which $\psi(i,q)$ is continuous in q across the equivalent classes.

3. Learning algorithms and analysis

In this section, we give a learning algorithm of reward-penalty type for MDPs with the transition matrices $q \in \mathbb{Q}^*$, by which the adaptive policy is constructed. For any $i \in S$ and $a \in A$, a sequence of stopping times $\{\sigma^n(i,a)\}_{n=0}^\infty$ will be defined as follows.

$$\sigma^0(i,a) = 0, \sigma^n(i,a) = \inf\{t | t > \sigma^{n-1}(i,a), X_t = i, \Delta_t = a\} \ (n \geqq 1). \tag{3.1}$$

Let $W := \bigcap_{(i,a) \in S \times A} W(i,a)$, where $W(i,a) = \bigcap_{n=1}^\infty \{\sigma^n(i,a) < \infty\}$. We note that $\omega \in W$ means that for any $(i,a) \in S \times A$ the event $\{X_t(\omega) = i, \Delta_t(\omega) = a\}$ happens in infinitely many stages.

The following is an extension of Lemma 1 in Kurano (1983) to the communicating case.

Lemma 3.1 *Let a policy $\pi = (\pi_0, \pi_1, \ldots)$ and a decreasing sequence of positive numbers $\{\varepsilon_t\}_{t=0}^{\infty}$ satisfy that*

(i) *for each $t \geqq 0, \pi_t(a|h_t) \geqq \varepsilon_t$ with $a \in A$ and $h_t = (x_0, a_0, \ldots, x_t) \in H_t$,*

(ii) $\displaystyle\sum_{t=0}^{\infty} \varepsilon_t^N = \infty.$

Then, $P_\pi(W \,\big|\, X_0 = i, q) = 1$ for all $q \in \mathbb{Q}^*$ and $i \in S$.

Proof. For notation simplicity, for any fixed $q \in \mathbb{Q}^*$ we put $P(\cdot) = P_\pi(\cdot|X_0 = i, q)$. From the definition of the communicating MDPs and (i) in Lemma 3.1, we have that there exists $\delta > 0$ such that

$$P(X_t = i, \Delta_t = a \text{ for some } t \text{ with } n \leqq t \leqq n+N|H_n) \geqq \delta\varepsilon_{n+N}^N \tag{3.2}$$

for any $n \geqq 0$ and $i \in S, a \in A$. For any fixed $i \in S, a \in A$, let $B_t := B_t(i,a) := \{\sigma^n(i,a) = t \text{ for some } n \geqq 1\}$. Then, we observe that $W(i,a) = \limsup_{t\to\infty} B_t(i,a) = \Big(\liminf_{t\to\infty} B_t^c(i,a)\Big)^c$, so that it holds that

$$P(W(i,a)) = 1 - P(\liminf_{t\to\infty} B_t^c(i,a)). \tag{3.3}$$

For any positive integer L with $L > n$, let $l := \Big[(L-n+1)/N\Big] - 1$, where for a real number z, $[z]$ is the largest integer equal to or less than z. Then, we have from (3.3) and (ii) in Lemma 3.1 that

$$\begin{aligned}
&P\Big(\bigcap_{t=n}^{L} B_t^c(i,a)\Big) \\
&\leqq P\Big(\bigcap_{\alpha=0}^{l} \bigcap_{t=n+\alpha N}^{n+(\alpha+1)N-1} B_t^c(i,a)\Big) \\
&\leqq \Big\{1 - P\Big(\bigcup_{t=n}^{n+N-1} B_t(i,a)\Big)\Big\} \cdots \Big\{1 - P\Big(\bigcup_{t=n+lN}^{n+(l+1)N-1} B_t(i,a) \,\Big|\, \bigcap_{t=n}^{n+lN-1} B_t^c(i,a)\Big)\Big\} \\
&\leqq \Big(1 - \delta\varepsilon_{n+N-1}^N\Big) \cdots \Big(1 - \delta\varepsilon_{n+(l+1)N-1}^N\Big) \\
&\leqq \exp\{-\delta \sum_{i=1}^{l+1} \varepsilon_{n+iN-1}^N\} \to 0 \text{ as } L \to \infty,
\end{aligned}$$

which implies that $\lim_{L\to\infty} P\Big(\bigcap_{t=n}^{L} B_t^c(i,a)\Big) = P\Big(\bigcap_{t=n}^{\infty} B_t^c(i,a)\Big) = 0$ for all $n \geqq 1$. Thus, from (3.3) $P(W(i,a)) = 1$, which implies $P(W) = 1$. □

We note that a sequence $\{1/(t+1)^{1/N}\}_{t=0}^{\infty}$ satisfies (ii) of Lemma 3.1 as an example.

For each $i, j \in S$ and $a \in A$, let $N_n(i,j|a) = \sum_{t=0}^{n} I_{\{X_t=i, \Delta_t=a, X_{t+1}=j\}}$ and $N_n(i|a) = \sum_{t=0}^{n} I_{\{X_t=i, \Delta_t=a\}}$, where I_D is the indicator function of a set D. Let

$$q_{ij}^n(a) = \begin{cases} \dfrac{N_n(i,j|a)}{N_n(i|a)} & \text{if } N_n(i|a) > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Then, $q_{ij}^n = (q_{ij}^n(a))$ is the maximum likelihood estimator of the unknown transition matrices. For any given $q^0 = (q_{ij}^0(a)) \in \mathbb{Q}$, we define $\tilde{q}^n = (\tilde{q}_{ij}^n(a)) \in \mathbb{Q}$ by

$$\tilde{q}_{ij}^n(a) = \begin{cases} q_{ij}^n(a) & \text{if } N_n(i|a) > 0, \\ q_{ij}^0(a) & \text{otherwise.} \end{cases}$$

We consider the following iterative scheme which is a variant of the non-stationary value iteration scheme proposed by Federgruen/Schweitzer (1981):

$$\tilde{v}_0 = 0, \quad \tilde{v}_{n+1} = U_\tau\{\tilde{q}^n\}\tilde{v}_n \quad (n \geqq 0) \tag{3.4}$$

for any $\tau \in (0,1)$. For each $i \in S$ and $n (n \geqq 0)$, let $\tilde{a}_{n+1}(i)$ denote an action which maximizes the right-hand side of the second equation in (3.4). For any sequence $\{b_n\}_{n=0}^\infty$ of positive numbers with $b_0 = 1, 0 < b_{n+1} < 1$ and $b_n > b_{n+1}$ for all $n \geqq 0$, let ϕ be any strictly increasing function such that $\phi : [0,1] \to [0,1]$ and $\phi(b_n) = b_{n+1}$ for all $n \geqq 0$.

Here, we define a learning algorithm based on $\tilde{a}_{n+1}$ and ϕ. For each $n (n \geqq 0)$, letting $\tilde{\pi}_n^\tau(a|i) = P(\Delta_n = a | X_0, \Delta_0, \ldots, X_n = i)$ we propose to update $\tilde{\pi}_n^\tau$ as follows: if $\tilde{a}_{n+1}(i) = a_i$ for each $i \in S$,

$$\begin{array}{ll} \tilde{\pi}_{n+1}^\tau(a_i|i) & = 1 - \sum_{a \neq a_i} \phi(\tilde{\pi}_n^\tau(a|i)), \\ \tilde{\pi}_{n+1}^\tau(a|i) & = \phi(\tilde{\pi}_n^\tau(a|i)) \ (a \neq a_i). \end{array} \tag{3.5}$$

In (3.5), the probability of choosing the action a_i at the next stage increases and that of choosing one of the other actions decreases, such that the algorithm (3.5) is a learning algorithm of the reward-penalty type (cf. Lakshmivarahan (1981), Meybodi/Lakshmivarahan (1982), Sutton and Barto (1998)). Note that given $\tilde{\pi}_0^\tau, \tilde{\pi}^\tau = (\tilde{\pi}_0^\tau, \tilde{\pi}_1^\tau, \ldots) \in \Pi$ and $\tilde{\pi}_n^\tau$ $(n \geqq 1)$ is successively determined by (3.4) and (3.5).

We need the following condition.

Condition A.

(i) $b_n \to 0$ as $n \to \infty$ and $\sum_{n=0}^\infty b_n^N = \infty$,

(ii) $\tilde{\pi}_0^\tau(a|i) > 0$ for all $i \in S, a \in A$.

Under this condition, the following lemma is proved similarly as Lemma 3 and 4 in Kurano (1987), so the proof is omitted.

Lemma 3.2 *Let $q \in \mathbb{Q}^*$. Then, under Condition A, the following (i)–(iii) holds with $P_{\tilde{\pi}^\tau}(\cdot|X_0 = i, q)$-a.s.:*

(i) $\tilde{q}^n \to q$ *as* $n \to \infty$,

(ii) $\tilde{v}_n(i) \to v_\tau(i,q)$ *as* $n \to \infty$,

(iii) $\tilde{\pi}_n^\tau(A_\tau^*(i|q)|H_n, X_n = i) \to 1$ *as* $n \to \infty$,

where $A_\tau^*(i|q)$ is the set of all actions which maximize the right-hand side of (1.6).

Let ${}^\tau\mathbb{Q}^* := \{q^{\tau,\mu} | \mu \in P(S) \text{ and } q \in \mathbb{Q}^*\}$, where $q^{\tau,\mu}$ is defined in (2.5). Then, observing the discussion in Section 2 and ${}^\tau\mathbb{Q}^* \subset \mathbb{Q}^*$, from Lemma 3.2 we find that the results in Kurano (1987) can be applicable to the class of perturbed transition matrices ${}^\tau\mathbb{Q}^*$. So, we have the following.

Theorem 3.1 *Under Condition A, $\tilde{\pi}^{\tau}$ is adaptively optimal for ${}^{\tau}\mathbb{Q}^{*}$.*

Here we can state the following theorem for the communicating case.

Theorem 3.2 *Under Condition A, a sequence $\{\tilde{\pi}^{\tau_n}\}_{n=1}^{\infty}$ with $\tau_n \to 0$ as $n \to \infty$ is an asymptotic sequence of adaptive policies with nearly optimal properties for $\mathbb{Q}^{*}$.*

Proof. Let $q \in \mathbb{Q}^{*}$. For each $t \geqq 0$, let

$$\begin{aligned}\tilde{\delta}_t &:= (1-\tau)\tilde{v}_t(X_t) - \{r(X_t, \Delta_t) + (1-\tau)\tilde{v}_t(X_{t+1})\} \text{ and}\\ \delta_t(j) &:= E_{\tilde{\pi}^{\tau}}(\tilde{\delta}_t | H_t, X_t = j, q).\end{aligned}$$

Then, by the stability theorem (cf. Loève (1963)), we get

$$\lim_{T\to\infty} \frac{1}{T+1}\sum_{t=0}^{T}\{\tilde{\delta}_t - \delta_t(X_t)\} = 0, \ P_{\tilde{\pi}^{\tau}}(\cdot | X_0 = i, q)\text{-a.s.} \tag{3.6}$$

On the other hand, it holds that

$$\delta_t(j) = \tilde{v}_t(j) - \sum_{a\in A}\Big\{r(j,a) + (1-\tau)\sum_{k\in S} q_{jk}(a)\tilde{v}_t(k)\Big\}\tilde{\pi}_t^{\tau}(a|j) - \tau\tilde{v}_t(j).$$

So, by (ii) and (iii) of Lemma 3.2,

$$\lim_{t\to\infty}\delta_t(j) = -\tau v_{\tau}(j,q), \quad P_{\tilde{\pi}^{\tau}}(\cdot | X_0 = i, q)\text{-a.s.}$$

Thus, from (3.6) it holds that

$$\begin{aligned}\min_{i\in S}\{-\tau v_{\tau}(i,q)\} &\leqq \liminf_{T\to\infty}\frac{1}{T+1}\sum_{t=0}^{T}\tilde{\delta}_t\\ &\leqq \limsup_{T\to\infty}\frac{1}{T+1}\sum_{t=0}^{T}\tilde{\delta}_t \leqq \max_{i\in S}\{-\tau v_{\tau}(i,q)\}.\end{aligned} \tag{3.7}$$

However, we have

$$\sum_{t=0}^{T}\tilde{\delta}_t = -\sum_{t=0}^{T} r(X_t, \Delta_t) + (1-\tau)\sum_{t=0}^{T}(\tilde{v}_t(X_t) - \tilde{v}_t(X_{t+1}))$$

by the definition. The second term in the right-hand-side is rewritten as

$$(1-\tau)\Big\{\tilde{v}_0(X_0) + \sum_{t=1}^{T}(\tilde{v}_t(X_t) - \tilde{v}_{t-1}(X_t)) - \tilde{v}_T(X_{T+1})\Big\}$$

so by (ii) of Lemma 3.2,

$$\limsup_{T\to\infty}\Big(\liminf_{T\to\infty}\Big)\frac{1}{T+1}\sum_{t=0}^{T}\tilde{\delta}_t = \limsup_{T\to\infty}\Big(\liminf_{T\to\infty}\Big)\Big\{-\frac{1}{T+1}\sum_{t=0}^{T} r(X_t, \Delta_t)\Big\}$$

respectively. Thus, applying Fatou's Lemma, from (3.7) we get

$$\min_{i\in S} \tau v_\tau(i,q) \leqq \psi(i,q|\tilde{\pi}^\tau) \leqq \max_{i\in S} \tau v_\tau(i,q). \tag{3.8}$$

By Lemma 1.2 and Theorem 2.1 (i), $\lim_{\tau\to 0} \tau v_\tau(i,q) = \psi(q)$, which implies from (3.8) that $\psi(i,q|\tilde{\pi}^\tau) \to \psi(q)$ as $\tau \to 0$. This completes the proof. □

We summarize our learning algorithm for constructing adaptive policy $\tilde{\pi}^\tau$ as follows:

Step 1. Set $n = 0$. Specify τ $(0 < \tau < 1)$. Choose $\tilde{\pi}_0^\tau(a|i)$ which satisfies (ii) in Condition A. Select $q_{ij}^0(a)$ arbitrarily and define $\tilde{v}_0(i) = 0$ $(i \in S)$.

Step 2. If $X_n = i$, choose an action $a_i \in A(i)$ from the decision rule $\tilde{\pi}_n^\tau$. Observe the next state $X_{n+1} = j$ and calculate $N_n(i,j|a)$ and $N_n(i|a)$.

Step 3. Set

$$\tilde{q}_{ij}^n(a) = \begin{cases} N_n(i,j|a)/N_n(i|a) & \text{if } N_n(i|a) > 0 \\ q_{ij}^0(a) & \text{otherwise.} \end{cases}$$

Step 4. Choose $\tilde{a}_{n+1}(i)$ which satisfies

$$\tilde{a}_i := \tilde{a}_{n+1}(i) \in \arg\max_{a\in A} \Big\{ r(i,a) + (1-\tau) \sum_{j\in S} \tilde{q}_{ij}^n(a) \tilde{v}_n(j) \Big\}.$$

Step 5. Update $\tilde{\pi}_{n+1}^\tau(i)$ $(i \in S)$ by

$$\tilde{\pi}_{n+1}^\tau(\alpha|i) = \phi\left(\tilde{\pi}_n^\tau(\alpha|i)\right) \ (\alpha \neq \tilde{a}_i), \text{ and } \tilde{\pi}_{n+1}^\tau(\tilde{a}_i|i) = 1 - \sum_{\alpha\neq\tilde{a}_i} \phi\left(\tilde{\pi}_n^\tau(\alpha|i)\right).$$

Step 6. Set $n = n + 1$ and return to Step 2.

Figure 3.1: Learning Algorithm for vanishing rate τ.

4. A numerical experiment

In this section, we give a simulation result for learning algorithm in Section 3.

Consider the three state MDPs with $S = \{1,2,3\}$ and $A = \{1,2\}$, whose transition matrices are parameterized with $0 < p_1, q_1, p_2, q_2 < 1$ and reward function $r(i,a)$ $(i \in S, a \in A)$ are given in Table 4.1 and Figure 4.1.

state	action	parameterized transition matrices			reward
i	a	$j=1$	$j=2$	$j=3$	$r(i,a)$
1	1	p_1	$1-p_1$	0	3 0
	2	$1-p_2$	p_2	0	2.5
2	1	0	q_1	$1-q_1$	2 0
	2	$1-q_2$	q_2	0	1.5
3	1	0	0	1	1 0
	2	0	1	0	0.5

Table 4.1: Data of simulated MDPs.

From this table and figure, we observe that the transition matrix q has a property of communicating provided that $0 < p_1, q_1, p_2, q_2 < 1$. We denote by $\tilde{\psi}_n$ the average

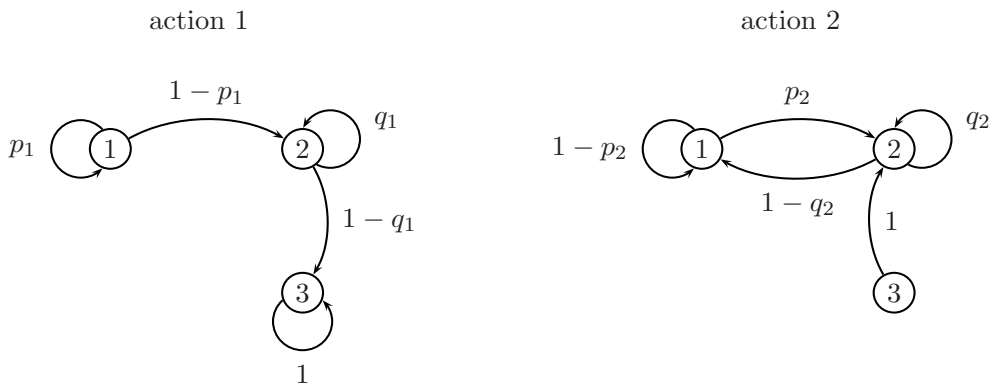

Figure 4.1: Transition diagrams parameterized with $0 < p_1, q_1, p_2, q_2 < 1$.

present value until n-th time, which is defined by

$$\tilde{\psi}_n = \frac{1}{n}\sum_{t=0}^{n-1} r(X_t, \Delta_t) \quad (n \geqq 1).$$

To calculate the quantity explicitly, we set $\tilde{\pi}_0^\tau(\cdot|i) = (\frac{1}{2}, \frac{1}{2})$ for each $i \in S$ and q_0 with $p_1 = \frac{2}{5}, q_1 = \frac{1}{2}, p_2 = \frac{3}{10}, q_2 = \frac{3}{10}$, i.e.,

$$q^0 = (q_{ij}^0(a)) = \Big\{(q_{ij}^0(1)) = \begin{pmatrix} \frac{2}{5} & \frac{3}{5} & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 \end{pmatrix}, (q_{ij}^0(2)) = \begin{pmatrix} \frac{7}{10} & \frac{3}{10} & 0 \\ \frac{7}{10} & \frac{3}{10} & 0 \\ 0 & 1 & 0 \end{pmatrix}\Big\}.$$

We use a strictly increasing function ϕ such that

$$\phi(x) = \Big(\frac{x^N}{1+x^N}\Big)^{1/N}$$

where N denotes the number of states in S. Let $\{b_n\}$ be such that $b_0 = 1$ and $b_n = n^{-1/N}$ $(n \geq 1)$. It is easily checked that the property (i) in Condition A is satisfied with $b_{n+1} = \phi(b_n)$ $(n \geqq 0)$.

Now, we make numerical experiments with the true transition matrices whose parameters are given by $p_1 = p_2 = \frac{1}{3}, q_1 = q_2 = \frac{2}{5}$, i.e.,

$$q = (q_{ij}(a)) = \Big\{(q_{ij}(1)) = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 \\ 0 & \frac{2}{5} & \frac{3}{5} \\ 0 & 0 & 1 \end{pmatrix}, (q_{ij}(2)) = \begin{pmatrix} \frac{2}{3} & \frac{1}{3} & 0 \\ \frac{3}{5} & \frac{2}{5} & 0 \\ 0 & 1 & 0 \end{pmatrix}\Big\}.$$

values	τ \ n	10^3	5×10^3	10^4	5×10^4	10^5	10^6
$\tilde{\psi}_n$	0.5	2.1104	2.1437	2.1569	2.1801	2.1876	2.2002
	0.2	2.1214	2.1468	2.1585	2.1805	2.1878	2.2002
	0.1	2.1224	2.1470	2.1586	2.1805	2.1878	2.2002
	0.01	2.1184	2.1462	2.1581	2.1804	2.1878	2.2002
decision rules	τ \ n	10^3	5×10^3	10^4	5×10^4	10^5	10^6
$\tilde{\pi}_n^\tau(1\|1)$	0.5	0.9003	0.9416	0.9536	0.9729	0.9785	0.9900
	0.2	0.8980	0.9413	0.9535	0.9728	0.9785	0.9900
	0.1	0.8983	0.9413	0.9535	0.9728	0.9785	0.9900
	0.01	0.8937	0.9409	0.9533	0.9728	0.9784	0.9900
$\tilde{\pi}_n^\tau(2\|2)$	0.5	0.8996	0.9415	0.9536	0.9729	0.9785	0.9900
	0.2	0.9002	0.9415	0.9536	0.9729	0.9785	0.9900
	0.1	0.9002	0.9415	0.9536	0.9729	0.9785	0.9900
	0.01	0.9002	0.9415	0.9536	0.9729	0.9785	0.9900
$\tilde{\pi}_n^\tau(2\|3)$	0.5	0.9002	0.9415	0.9536	0.9729	0.9785	0.9900
	0.2	0.9002	0.9415	0.9536	0.9729	0.9785	0.9900
	0.1	0.9002	0.9415	0.9536	0.9729	0.9785	0.9900
	0.01	0.9002	0.9415	0.9536	0.9729	0.9785	0.9900

Table 4.2: The simulation value of $\tilde{\psi}_n$ and $\tilde{\pi}_n^\tau$ for each $\tau = 0.5, 0.2, 0.1, 0.01$.

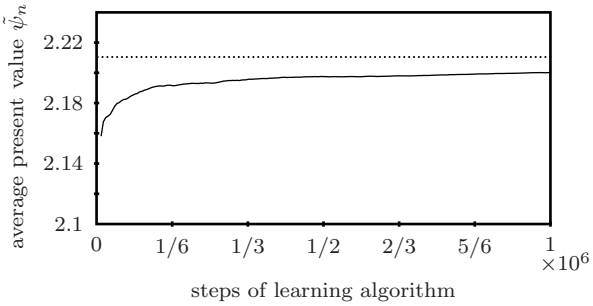

Figure 4.2: The trajectories of $\tilde{\psi}_n(\tau = 0.01)$. The dotted line means the optimal value of average reward.

Considering the optimal average reward $\psi(q) = 42/19 \approx 2.2105$ and the q-optimal stationary policy f^* with $f^*(1) = 1, f^*(2) = f^*(3) = 2$, it is seen that $\tilde{\psi}_n \to \psi(q) = 42/19$ and $\tilde{\pi}_n^\tau(1|1)$, $\tilde{\pi}_n^\tau(2|2)$, $\tilde{\pi}_n^\tau(2|3) \to 1$ as $n \to \infty$ hold from the above Table 4.2 and Figure 4.2. The results of the above simulation show that the learning algorithm is practically effective for the communicating class of transition matrices.

References

Bather, J. (1973). Optimal decision procedures for finite Markov chains. II. Communicating systems. *Adv. in Appl. Probab.*, 5:521–540.

Bertsekas, D. P. and Tsitsiklis, J. H. (1996). *Neuro-Dynamic Programming.* Athena Scientific.

Billingsley, P. (1961). *Statistical inference for Markov processes.* The University of Chicago Press.

Federgruen, A. and Schweitzer, P. J. (1981). Nonstationary Markov decision problems with converging parameters. *J. Optim. Theory Appl.*, 34(2):207–241.

Hernández-Lerma, O. (1989). *Adaptive Markov control processes.* Springer-Verlag.

Hernández-Lerma, O. and Marcus, S. I. (1985). Adaptive control of discounted Markov decision chains. *J. Optim. Theory Appl.*, 46(2):227–235.

Kemeny, J. G. and Snell, J. L. (1960). *Finite Markov chains.* D. Van Nostrand.

Kurano, M. (1972). Discrete-time Markovian decision processes with an unknown parameter. Average return criterion. *J. Operations Res. Soc. Japan*, 15:67–76.

Kurano, M. (1983). Adaptive policies in Markov decision processes with uncertain transition matrices. *J. Inform. Optim. Sci.*, 4(1):21–40.

Kurano, M. (1987). Learning algorithms for Markov decision processes. *J. Appl. Probab.*, 24(1):270–276.

Lakshmivarahan, S. (1981). *Learning algorithms.* Springer-Verlag.

Leizarowitz, A. (2003). An algorithm to identify and compute average optimal policies in multichain Markov decision processes. *Math. Oper. Res.*, 28(3):553–586.

Loève, M. (1963). *Probability theory.* D. Van Nostrand.

Mandl, P. (1974). Estimation and control in Markov chains. *Adv. in Appl. Probab.*, 6:40–60.

Martin, J. J. (1967). *Bayesian decision problems and Markov chains.* John Wiley & Sons Inc.

Meybodi, M. R. and Lakshmivarahan, S. (1982). ε-optimality of a general class of learning algorithms. *Inform. Sci.*, 28(1):1–20.

Puterman, M. L. (1994). *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons Inc.

Ross, S. M. (1970). *Applied probability models with optimization applications.* Holden-Day.

Schweitzer, P. J. (1968). Perturbation theory and finite Markov chains. *J. Appl. Probability*, 5:401–413.

Solan, E. (2003). Continuity of the value of competitive Markov decision processes. *J. Theoret. Probab.*, 16(4):831–845 (2004).

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introducution.* MIT Press.

van Hee, K. M. (1978). *Bayesian control of Markov chains.* Mathematisch Centrum.

Received June 7, 2006
Revised February 25, 2007