数理解析研究所講究録
第 1589 巻 2008 年 110-119

A pattern-matrix learning algorithm for adaptive MDPs: The regularly communicating case

宮崎大学・教育文化学部 伊喜 哲一郎 (Tetsuichiro IKI)
Faculty of Education and Culture, Miyazaki University
弓削商船高等専門学校・総合教育科 堀口 正之 (Masayuki HORIGUCHI)
General Education, Yuge National College of Maritime Technology
千葉大学・理学部 安田 正實 (Masami YASUDA)
Faculty of Science, Chiba University
千葉大学・教育学部 蔵野 正美 (Masami KURANO)
Faculty of Education, Chiba University

Abstract

In this note, as a sequel to our previous work[7], we are concerned with adaptive models for uncertain Markov decision processes with regularly communicating structure where the state space is decomposed into a single communicating class and a absolutely transient class.

We give a pattern-matrix learning algorithm which finds the regularly communicating structure, by which an asymptotic sequence of adaptive properties with nearly average-optimal properties is constructed. A numerical experiment is given.

Keywords: adaptive Markov decision processes, pattern-matrix learning algorithm, average-optimal adaptive policy, regularly communicating case.

1 Introduction and notation

In our previous work[7], we considered the adaptive Markov decision processes(MDPs) in which the state space is a single communicating class and constructed an average-optimal adaptive policy of reward-penalty types(cf. [9, 10]) by applying the perturbation theory(cf. [16]).

In this note, as a sequel to [7], we are concerned with adaptive models for uncertain MDPs with regularly communicating structure where the state space is assumed to be decomposed into a single communicating class and a transient class(cf. [1, 6, 11]). In this case, the corresponding adaptive policy will be compelled to learn the pattern of the structure.

Here, we give a pattern-matrix learning algorithm for regularly communicating structure, by which an asymptotic sequence of adaptive properties with nearly average-optimal properties is constructed by extending the results of [7].

For general discussions of adaptive MDPs, refer to [4, 5, 12, 13, 18] and for an approach by the neuro-dynamic programming refer to [2, 8, 17].

In the reminder of this section, we formulate the adaptive MDPs with uncertain transition matrices.

Consider a controlled dynamic system with finite state space $S = \{1, 2, \ldots, N\}$, containing $N < \infty$ elements. For each $i \in S$, the finite set $A(i)$ denotes the set of available actions at state i. Let $\mathbb{Q}$ denote the parameter space of unknown transition matrices, i.e.,

$$\mathbb{Q} = \{q = (q_{ij}(a)) | q_{ij}(a) \geqq 0, \sum_{j \in S} q_{ij}(a) = 1 \text{ for } i, j \in S \text{ and } a \in A(i)\}. \tag{1.1}$$

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:

$$\psi(i, q|\pi) = \liminf_{T \to \infty} \frac{1}{T+1} E_\pi \left(\sum_{t=0}^{T} r(X_t, \Delta_t) \,\middle|\, X_0 = i, q \right) \tag{1.2}$$

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

$$\psi(i, q) = \sup_{\pi \in \Pi} \psi(i, q|\pi), \tag{1.3}$$

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 $q \in \mathcal{D}$.

Let $q \in \mathbb{Q}$. A subset $E \subset S$ is called a communicating class for q if

(i) for any $i, j \in E$, there exists a path in E from i to j with positive probability, rewritten by "$i \to j$", i.e., it holds that

$$q_{i_1 i_2}(a_1) q_{i_2 i_3}(a_2) \cdots q_{i_{l-1} i_l}(a_{l-1}) > 0 \tag{1.4}$$

for some $\{i_1 = i, i_2, \ldots, i_l = j\} \subset E$ and $a_k \in A(i_k)$ and $2 \leqq l \leqq N$, and

(ii) E is closed, i.e., $\sum_{j \in E} q_{ij}(a) = 1$ for $i \in E, a \in A(i)$.

The transition matrix $q \in \mathbb{Q}$ is said to be regularly communicating if there exists an $\bar{E} \subsetneqq S$ such that

(i) $\bar{E}$ is a communicating class for q and

(ii) $T = S - \bar{E}$ is an absolutely transient class, i.e.,

$$P_\pi(X_t \in \bar{E} \text{ for some } t \geqq 1 | X_0 \in T) = 1 \tag{1.5}$$

for all $\pi \in \Pi$

For a regularly communicating $q \in \mathbb{Q}$, this corresponding communicating class $\bar{E}$ will be denoted by $\bar{E}(q)$ depending on $q \in \mathbb{Q}$. For any $i_0 \in S$, we denote by $\mathbb{Q}^*(i_0)$ the set of regularly communicating $q \in \mathbb{Q}$ with $i_0 \in \bar{E}(q)$.

Let $n(D)$ denotes the number of elements in a set D. For any $q \in \mathbb{Q}^*(i_0)$, the pattern-matrix $M(q)$ (cf. [6]) corresponding with q is generally represented as follows:

$$M(q) = \left(\begin{array}{c:c} E & O \\ \hdashline R & K \end{array}\right)$$

where E is an $n(\bar{E}(q)) \times n(\bar{E}(q))$-matrix and R is an $n(S - \bar{E}(q)) \times n(\bar{E}(q))$-matrix whose elements of both E and R are all 1 and that $i \to j$ means that the (i,j) element of $M(q)$ is 1.

The adaptive policy for $q \in Q^*(i_0)$ will be necessary to find the pattern-matrix $M(q)$, whose algorithm will be called the pattern-matrix learning one.

The sequence of policies $\{\tilde{\pi}^n\}_{n=0}^{\infty} \subset \Pi$ is called an asymptotic sequence of adaptive policies with nearly optimal properties for $\mathcal{D} \subset \mathbb{Q}$ and $E \subset S$ if

$$\lim_{n\to\infty} \psi(i, q|\tilde{\pi}^n) = \psi(i, q) \tag{1.6}$$

for all $q \in \mathcal{D}$ and $i \in E$.

In [9], an adaptively optimal policy for

$$\mathbb{Q}^+ := \{q = (q_{ij}(a)) \in \mathbb{Q} | q_{ij}(a) > 0 \text{ for all } i, j \in S \text{ and } a \in A(i)\}, \tag{1.7}$$

was constructed by applying the value iteration and policy improvement algorithm (cf. [3]) which was extensively applied to the communicating case of multi-chain MDPs in Iki et. al. [7].

In this note, using the method of pattern-matrix learning we will construct an asymptotic sequence of adaptive policies with nearly optimal properties for $\mathbb{Q}^*(i_0)$ with $i_0 \in S$, which is thought of as a wider class for uncertain MDPs than the communicating case treated in [7]. In order to treat with the regularly communicating case with $q \in \mathbb{Q}^*(i_0)$, 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

$$v_\tau(i, q|\pi) = E_\pi\left(\sum_{t=0}^{\infty}(1-\tau)^t r(X_t, \Delta_t)|X_0 = i, q\right) \tag{1.8}$$

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.

Let $B(S)$ be the set of all functions on S. 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

$$U_\tau\{q\}u(i) = \max_{a\in A}\left\{r(i, a) + (1-\tau)\sum_{j\in S} q_{ij}(a)u(j)\right\} \tag{1.9}$$

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

Lemma 1.1 ([14, 15]). *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.,*

$$v_\tau = U_\tau\{q\}v_\tau, \tag{1.10}$$

(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 (1.10).*

In Section 2, some elementary lemmas are given which show the effectiveness of pattern-matrix leaning algorithm developed in the sequel. Section 3 is devoted to the construction of adaptive policies with nearly average-optimal properties for $\mathbb{Q}^*(i_0)$. A numerical experiment is given in Section 4.

2 Preliminary lemmas

In this section, several lemmas are given which are used in Section 3.

Let $i_0 \in S$. For any $q \in \mathbb{Q}^*(i_0)$ and $E \subsetneqq \bar{E}(q)$, we define the sequence $J_k(E)$ $(k = 1,2,\ldots)$ iteratively by

$$J_1(E) = \{i \in E| \sum_{j\in\bar{E}(q)-E} q_{ij}(a) > 0 \text{ for some } a \in A(i)\}$$

and (2.1)

$$J_k(E) = \{i \in E - \bigcup_{l=1}^{k-1} J_l(E)| \sum_{j\in J_{k-1}(E)} q_{ij}(a) > 0 \text{ for some } a \in A(i)\} \ (k \geqq 2).$$

Letting $K(\bar{E}(q)) = \{(i,a,j)|p_{ij}(a) > 0, i,j \in \bar{E}(q) \text{ and } a \in A(i)\}$, put $\delta := \min p_{ij}(a)$ where the minimum is taken over $(i,a,j) \in K(\bar{E}(q))$. Then, from the definition of communicating class $\bar{E}(q)$, the following can be easily shown.

Lemma 2.1. *For any* $q \in \mathbb{Q}^*(i_0)$ *with* $i_0 \in S$ *and* $E \subsetneqq \bar{E}(q)$, *there exists* $l(E)$ $(1 \leqq l(E) \leqq N)$ *for which* $J_k(E) \neq \emptyset$ $(k = 1,2,\ldots,l(E))$ *and* $J_{l(E)+1}(E) = \emptyset$.

Lemma 2.2. *Let* $q \in \mathbb{Q}^*(i_0)$ *with* $i_0 \in S$. *Let a policy* $\bar{\pi} = (\bar{\pi}_0, \bar{\pi}_1, \ldots)$ *and a decreasing sequence of positive numbers* $\{\varepsilon_t\}_{t=0}^\infty$ *satisfy that for each* $t \geqq 0$ $\bar{\pi}_t(a|h_t) \geqq \varepsilon_t$ *with* $a \in A(x_t)$ *and* $h_t = (x_0, a_0, x_1, \ldots, x_t) \in H_t$. *Then, it holds that for any* $E \subsetneqq \bar{E}(q)$,

$$P_{\bar{\pi}}(X_{t+l} \in \bar{E}(q) - E \text{ for some } l(1 \leqq l \leqq N)|X_t \in E) \geqq (\delta\varepsilon_{t+N})^N. \quad (2.2)$$

Proof. By Lemma 2.1, it holds that

$$\begin{aligned}\text{the left-hand side of (2.2)} &\geqq (\varepsilon_t\delta)(\varepsilon_{t+1}\delta)\cdots(\varepsilon_{t+l(E)}\delta)\\ &\geqq (\delta\varepsilon_{t+N})^N,\end{aligned}$$

which completes the proof.∎

For $q \in \mathbb{Q}^*(i_0)$ with $i_0 \in S$, a sequence of stopping times $\{\sigma_t\}$ and subsets $\{E_{\sigma_t}\} \subset \bar{E}(q)$ will be defined as follows:

$E_0 := \{i_0\}, T_0 := \bar{E}(q) - E_0, \sigma_1 := \min\{t|X_t \in T_0, t > 0\}$,
$E_{\sigma_1} = E_0 \cup \{X_{\sigma_1}\}, T_{\sigma_1} := \bar{E}(q) - E_{\sigma_1}$,
and iteratively for $n = 2,3,\ldots,$ (2.3)
$\sigma_n := \min\{t|X_t \in T_{\sigma_{n-1}}, t > \sigma_{n-1}\}, E_{\sigma_n} = E_{\sigma_{n-1}} \cup \{X_{\sigma_n}\}, T_{\sigma_n} = \bar{E}(q) - E_{\sigma_n}$,
where $\min\emptyset = \infty$.

For any $E \subset \bar{E}(q)$, let $\bar{n}(E) = \min\{n \geqq 1|E_{\sigma_n} = \bar{E}(q)\}$. If $\bar{n}(E) < \infty$, we can find the pattern-matrix $M(q)$. Here, we have the following.

Lemma 2.3. *Let $q \in \mathbb{Q}^*(i_0)$ with $i_0 \in S$ and $\tilde{\pi}$ satisfy condition in Lemma 2.2 with $\sum_{t=0}^{\infty} \varepsilon_t^N = \infty$. Then, for any $E \subsetneqq \bar{E}(q)$ it holds that*

(i) $P_{\tilde{\pi}}(\bar{n}(E) < \infty | X_0 = i_0, q) = 1$, *and*

(ii) for any $k \leqq \bar{n}(E), P_{\tilde{\pi}}(\sigma_k < \infty | X_0 = i_0, q) = 1$.

Proof. For any $E \subsetneqq \bar{E}(q)$, from Lemma 2.2 and $\sum_{t=0}^{\infty} \varepsilon_t^N = \infty$ it follows that

$$P_{\tilde{\pi}}(X_{t+l} \in E \text{ for all } l \geqq 1 | X_t \in E, q) \leqq \prod_{l=1}^{\infty}(1 - \delta^N \varepsilon_{t+lN}^N) \leqq e^{-\delta^N} \sum_{l=1}^{\infty} \varepsilon_{t+lN}^N = 0. \tag{2.4}$$

So, taking $E = E_0$ in (2.4), we have

$$\begin{aligned} P_{\tilde{\pi}}(\sigma_1 < \infty | X_0 \in E_0, q) &= 1 - P_{\tilde{\pi}}(\sigma_1 = \infty | X_0 \in E_0, q) \\ &= 1 - P_{\tilde{\pi}}(X_t \in E_0 \text{ for all } t \geqq 1 | X_0 \in E_0, q) \\ &= 1. \end{aligned}$$

For (ii), inductively on k ($k = 2, 3, \ldots$), if $E_{\sigma_{k-1}} \subsetneqq \bar{E}(q)$, we have from (2.4) that

$$\begin{aligned} &P_{\tilde{\pi}}(\sigma_k < \infty | X_0 \in E_0, q) \\ &= \sum_{l=1}^{\infty} P_{\tilde{\pi}}(\sigma_{k-1} = l | X_0 \in E_0, q) \cdot P_{\tilde{\pi}}(X_{t+l} \in \bar{E}(q) - E_l \text{ for some } 0 < t < \infty | X_l \in E_l, q) \\ &= \sum_{l=1}^{\infty} P_{\tilde{\pi}}(\sigma_{k-1} = l | X_0 \in E_0, q) \\ &= P_{\tilde{\pi}}(\sigma_{k-1} < \infty | X_0 \in E_0, q) \\ &= 1. \end{aligned} \tag{2.5}$$

Obviously, (i) follows from (ii), which completes the proof.∎

We note that a sequence $\{(1+t)^{-N}\}_{t=0}^{\infty}$ satisfies Assumption concerning $\{\varepsilon_t\}_{t=0}^{\infty}$ given in Lemma 2.3.

3 Pattern-matrix learning algorithms

In this section, we give a pattern-matrix learning algorithm by which an asymptotic sequence of adaptive policies with nearly average-optimal properties for $\mathbb{Q}^*(i_0)$ with $i_0 \in S$ is given.

For any sequence $\{b_n\}_{n=0}^{\infty}$ of positive numbers with $b_0 = 1, 0 < b_n < 1$ and $b_n > b_{n+1}$ for all $n \geqq 1$, let ϕ be any strictly increasing function that $\phi : [0,1] \to [0,1]$ and $\phi(b_n) = b_{n+1}$ for all $n \geqq 0$.

Here, we consider the following iterative scheme called a pattern-matrix learning algorithm with $i_0 \in S, \{b_n\}$ and $\tau \in (0,1)$, denoted by $\mathbf{PMLA}(i_0, \{b_n\}, \tau)$.

$\underline{\mathbf{PMLA}(i_0, \{b_n\}, \tau)}$:

1. Set $E_0 = \{i_0\}, T_0 = S - E_0, \tilde{v}_0(i) = 0$ $(i \in E_0)$, $X_0 = i_0$ and $\tilde{\pi}_0^{\tau}(a|X_0) = n(A(i_0))^{-1}$ for $a \in A(i_0)$.

2. Suppose that $E_n \subset S, T_n = S - E_n$ and $\{\tilde{v}_n(i) : i \in E_n\}$ are given. Moreover, suppose that the n-th decision rule $\tilde{\pi}_n^{\tau}(a|i) = Prob.(\Delta_n = a|H_{n-1}, \Delta_{n-1}, X_n = i)$ $(i \in E_n, a \in A(i))$ are given, where $H_{n-1} = (X_0, \Delta_0, X_1, \ldots, X_{n-1})$ is a history until the $(n-1)$-th step.

3. Choose an action $\Delta_{n+1} \in A(X_n)$ from $\tilde{\pi}_n(\cdot|H_n)$. Then, according to the value of X_{n+1}, we put $E_{n+1} = E_n \cup \{X_{n+1}\}$ if $X_{n+1} \in T_n$ and $E_{n+1} = E_n$ if $X_n \in E_n$.
   Calculate $N_{n+1}(i,j|a) = \sum_{t=0}^{n} I_{\{X_t=i,\Delta_t=a,X_{t+1}=j\}}$ and $N_{n+1}(i|a) = \sum_{t=0}^{n} I_{\{X_t=i,\Delta_t=a\}}$ for $i,j \in E_{n+1}$ and $a \in A(i)$.
   Set $q^{n+1} = (q_{ij}^{n+1}(a))$ by
$$q_{ij}^{n+1}(a) = \begin{cases} \dfrac{N_{n+1}(i,j|a)}{N_{n+1}(i,a)} & \text{if } N_{n+1}(i|a) > 0, \\ q_j^0 & \text{otherwise,} \end{cases} \quad (i,j \in E_{n+1}, a \in A(i)) \tag{3.1}$$
   where $q^0 = (q_j^0 : j \in E_{n+1})$ is any distribution on E_{n+1} with $q_j^0 > 0$ for all $i \in E_{n+1}$.

4. For each $i \in E_{n+1}$, choose $\tilde{a}_{n+1}(i)$ which satisfies
$$\tilde{a}_{n+1}(i) \in \arg\max_{a \in A(i)}\{r(i,a) + (1-\tau)\sum_{j \in E_{n+1}} q_{ij}^{n+1}(a)\tilde{v}_n(j)\}$$
   and update $\tilde{\pi}_{n+1}^{\tau}(a|i) = Prob.(\Delta_{n+1} = a|H_n, \Delta_{n+1}, X_{n+1} = i)$ as follows:
$$\tilde{\pi}_{n+1}^{\tau}(a_i|i) = \begin{cases} 1 - \sum_{a \neq a_i} \phi(\tilde{\pi}_n^{\tau}(a|i)) & (a_i = \tilde{a}_{n+1}(i)) \\ \phi(\tilde{\pi}_n^{\tau}(a_i|i)) & (a_i \neq \tilde{a}_{n+1}(i)). \end{cases} \tag{3.2}$$
   Moreover, put $\tilde{v}_{n+1} = U_\tau\{q^{n+1}\}\tilde{v}_n$ on E_{n+1}.

5. Set $n \leftarrow n+1$ and return to step 3.

We need the following condition on $\{b_n\}$.

Condition (*)

$$b_n \to 0 \text{ as } n \to \infty \text{ and } \sum_{n=0}^{\infty} b_n^N = \infty. \tag{3.3}$$

The following theorem says that the policy $\tilde{\pi}^{\tau} = (\tilde{\pi}_0^{\tau}, \tilde{\pi}_1^{\tau}, \ldots)$ constructed by **PMLA(i_0, $\{b_n\}$, τ)** has nearly average-optimal properties for $\mathbb{Q}^*(i_0)$ when $\tau \to 0$.

Theorem 3.1. *Under condition (*), 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 average-optimal properties for $\mathbb{Q}^*(i_0)$.*

Proof. Under condition (*), the policy $\tilde{\pi}^{\tau} = (\tilde{\pi}_0^{\tau}, \tilde{\pi}_1^{\tau}, \ldots)$ constructed in **PMLA(i_0, $\{b_n\}$, τ)** satisfies assumptions in Lemma 2.3. So, by Lemma 2.3 we observe that **PMLA(i_0, $\{b_n\}$, τ)** finds the pattern $\bar{E}(q)$ with $P_{\tilde{\pi}^{\tau}}(\cdot|X_0 = i_0, q)$-probability 1, i.e.,
$$E_n = \bar{E}(q) \text{ for all } n \geqq \bar{n}(E_0),$$
where $\bar{n}(E_0)$ is given in Lemma 2.3.

Thus, a learning algorithm for communicating MDPs on $\bar{E}(q)$ for $q \in \mathbb{Q}(i_0)$, which was developed in [7] using the vanishing discount approach(Lemma 1.1), are applicable to the pattern-matrix learning case, which completes the proof.∎

4 A numerical experiment

In this section, we give a simulation result for pattern-matrix learning algorithm.

Consider the six-state MDPs with $S = \{1, 2, 3, 4, 5, 6\}$, where data for simulation and transition diagrams are given in Table 4.1. and Fig. 4.1.

Table 4.1: Data of simulated MDPs

state	action	transition probabilities $q_{ij}(a)$						reward
i	$a \in A(i)$	$j=1$	$j=2$	$j=3$	$j=4$	$j=5$	$j=6$	$r(i,a)$
1	1	0	$\frac{1}{2}$	0	$\frac{1}{2}$	0	0	9
	2	$\frac{1}{4}$	0	$\frac{1}{4}$	0	$\frac{1}{2}$	0	10
2	1	0	$\frac{1}{2}$	0	0	$\frac{1}{4}$	$\frac{1}{4}$	5
	2	0	0	1	0	0	0	2
3	1	0	0	$\frac{2}{5}$	0	$\frac{3}{5}$	0	7
	2	0	$\frac{1}{2}$	0	0	0	$\frac{1}{2}$	8
4	1	$\frac{1}{2}$	0	$\frac{1}{2}$	0	0	0	2
	2	0	$\frac{1}{4}$	0	$\frac{1}{4}$	$\frac{1}{2}$	0	12
5	1	0	$\frac{1}{4}$	0	0	$\frac{1}{2}$	$\frac{1}{4}$	6
	2	0	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0	2.5
	3	0	$\frac{1}{2}$	0	0	$\frac{1}{2}$	0	2.25
6	1	0	$\frac{1}{2}$	0	0	$\frac{1}{4}$	$\frac{1}{4}$	14
	2	0	0	$\frac{1}{2}$	0	0	$\frac{1}{2}$	8

We denote by $\tilde{\psi}_n$ the average present value until n-th time, 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 $E_0 = \{2\}$. We use a strictly increasing function ϕ such that

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

where N denotes the number of states in S.

The pattern matrix $M(q)$ and reordered matrix $\overline{M}$ corresponding to communicating states are easily computed, which are shown as follows.

$$M = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 \end{pmatrix} \end{array}, \quad \overline{M} = \begin{array}{c} \\ 2 \\ 3 \\ 5 \\ 6 \\ 1 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccccc} 2 & 3 & 5 & 6 & 1 & 4 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} \end{array}.$$

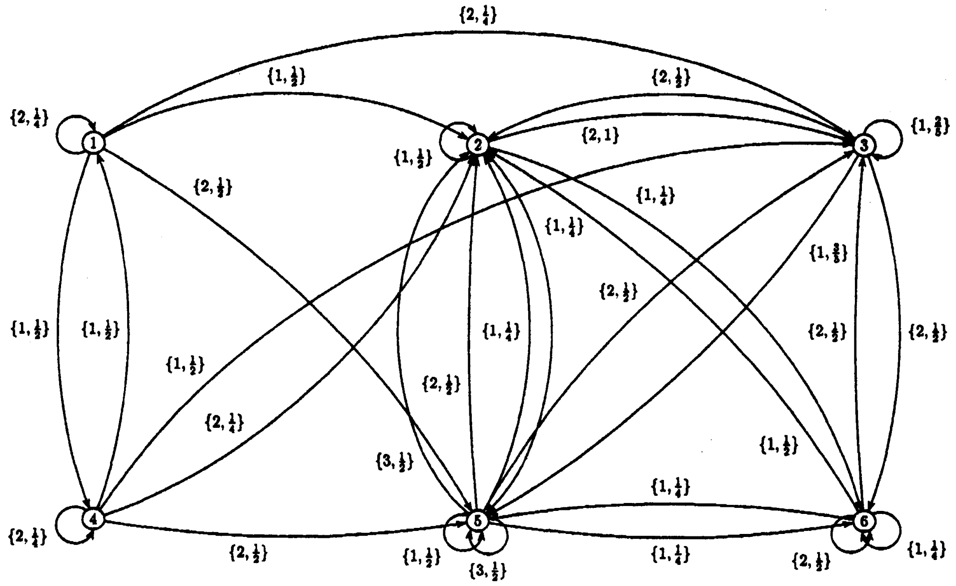

Figure 4.1: Transition diagrams of numerical experiment. The first quantity in brackets near the arc is action number and the second one is its transition probability if the action is chosen.

Now, we make numerical experiments with vanishing parameter $\tau = 0.1$ and 0.01 and show the results given in Table 4.2. and Fig. 4.2.

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

values	τ \ n	10^3	5×10^3	10^4	5×10^4	10^5	10^6	10^7
$\tilde{\psi}_n$	0.10	6.403347	6.672316	6.827892	6.965801	7.013102	7.158738	7.282986
	0.01	6.365634	6.651570	6.816118	6.963271	7.011827	7.158618	7.282973
decision	τ \ n	10^3	5×10^3	10^4	5×10^4	10^5	10^6	10^7
$\tilde{\pi}_n^\tau(1\|2)$	0.10	0.661198	0.755328	0.783315	0.835058	0.853137	0.899994	0.931870
	0.01	0.493097	0.749621	0.780978	0.834723	0.852989	0.899984	0.931870
$\tilde{\pi}_n^\tau(2\|3)$	0.10	0.685394	0.758424	0.784669	0.835262	0.853228	0.900001	0.931871
	0.01	0.685111	0.758380	0.784649	0.835259	0.853226	0.900000	0.931871
$\tilde{\pi}_n^\tau(1\|5)$	0.10	0.422566	0.527159	0.574120	0.671281	0.706794	0.800024	0.863743
	0.01	0.422566	0.527159	0.574120	0.671281	0.706794	0.800024	0.863743
$\tilde{\pi}_n^\tau(1\|6)$	0.10	0.686510	0.758602	0.784748	0.835274	0.853233	0.900001	0.931871
	0.01	0.686510	0.758602	0.784748	0.835274	0.853233	0.900001	0.931871

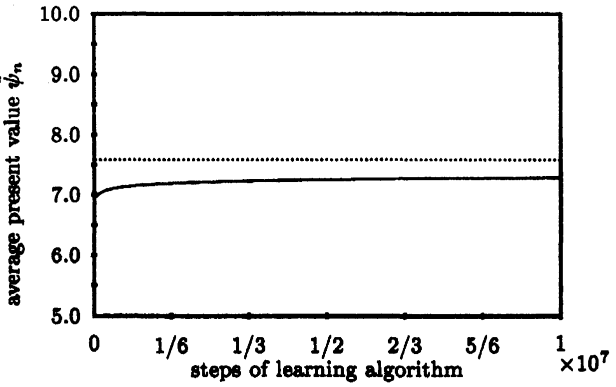

Figure 4.2: The trajectories of $\tilde{\psi}_n(\tau = 0.01)$. The dotted line means the optimal value of average reward in $\bar{E}(q)$.

Considering the optimal average reward $\psi(i, q) = 91/12 \approx 7.583(i \in \bar{E}(q))$ and the q-optimal stationary policy f^* for $\bar{E}(q)$ is $f^*(2) = 1, f^*(3) = 2, f^*(5) = 1, f^*(6) = 1$, it is seen that $\tilde{\psi}_n \to \psi(i, q) = 91/12$ and $\tilde{\pi}^\tau_n(1|1)$, $\tilde{\pi}^\tau_n(2|2)$, $\tilde{\pi}^\tau_n(2|3) \to 1$ as $n \to \infty$ hold from the above Table 4.2 and Fig. 4.2. The results of the above simulation show that the pattern-matrix learning algorithm is practically effective for the communicating class of transition matrices.

References

[1] John Bather. Optimal decision procedures for finite Markov chains. II. Communicating systems. *Advances in Appl. Probability*, 5:521–540, 1973.

[2] D.P. Bertsekas and J.H. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Massachusetts, Belmont, 1996.

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

[4] O. Hernández-Lerma. *Adaptive Markov control processes*, volume 79 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1989.

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

[6] T. Iki, M. Horiguchi, and M. Kurano. A structured pattern matrix algorithm for multichain markov decision processes. *Mathematical Methods of Operations Research,*, electronic version, 2007.

[7] T. Iki, M. Horiguchi, M. Yasuda, and M. Kurano. A learning algorithm for communicating markov decision processes with unknown transition matrices. *(to appear in Bulletin of Information and Cybernetics)*, 2007.

[8] T. Iki, M. Horiguchi, M. Yasuda, and M. Kurano. Temporal difference-based adaptive policies in neuro-dynamic programming. *Vicenc Torra, Yasuo Narukawa, Yuji Yoshida (Eds.), 4th International conference on Proceedings of Modeling Decisions for Artificial*

Intelligence (MDAI) 2007 (CD-ROM Proceedings, ISBN 978-84-00-08539-1), pages 112–122, 2007.

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

[10] S. Lakshmivarahan. *Learning algorithms.* Springer-Verlag, New York, 1981. Theory and applications.

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

[12] P. Mandl. Estimation and control in Markov chains. *Advances in Appl. Probability*, 6:40–60, 1974.

[13] J. J. Martin. *Bayesian decision problems and Markov chains.* Publications in Operations Research, No. 13. John Wiley & Sons Inc., New York, 1967.

[14] Martin L. Puterman. *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons Inc., New York, 1994. A Wiley-Interscience Publication.

[15] Sheldon M. Ross. *Applied probability models with optimization applications.* Holden-Day, San Francisco, Calif., 1970.

[16] Paul J. Schweitzer. Perturbation theory and finite Markov chains. *J. Appl. Probability*, 5:401–413, 1968.

[17] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction.* Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 1998.

[18] K. M. van Hee. *Bayesian control of Markov chains*, volume 95 of *Mathematical Centre Tracts.* Mathematisch Centrum, Amsterdam, 1978.