数理解析研究所講究録 1477 巻 2006 年 127-135

# A learning algorithm for communicating Markov decision processes with unknown transition matrices

宮崎大学・教育文化学部　伊喜　哲一郎 (Tetsuichiro IKI)
Faculty of Education and Culture, Miyazaki University
東京電機大学・情報環境学部　堀口　正之 (Masayuki HORIGUCHI)
School of Information Environment, Tokyo Denki University
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　藏野　正美 (Masami KURANO)
Faculty of Education, Chiba University

## Abstract

This study is concerned with finite Markov decision processes(MDPs) whose transition matrices are unknown but the state is observable exactly. We develop a learning algorithm of the reward-penalty type for the communicating case of multi-chain MDPs by which 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:** Adaptive policy, Markov decision processes, communicating case, average case, learning algorithm, reward-penalty type.

## 1 Introduction and notation

Markov decision processes(MDPs) whose transition probabilities are unknown to a decision maker have been investigated by many authors (cf. [4, 5, 8, 9, 14, 15]). Kurano[10] proposed a learning algorithms of the reward-penalty type(cf. [11, 16]) for the positive case where all elements of the true transition matrices of finite MDPs are known to be positive, by which adaptively optimal policy are constructed under the average expected reward criterion.

In this paper, applying the idea of [10] extensively to a large class of uncertain MDPs, we develop a learning algorithm for the communicating case of multi-chain MDPs 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 $\mathbb{Q} = \{q | q = (q_{ij}(a) : i, j \in S, a \in A), q_{ij}(a) \geqq 0, \sum_{j \in S} q_{ij}(a) = 1 \text{ for } i, j \in S, a \in A\}$.

The sample space is the product space $\Omega = (S \times A)^\infty$ such that the projections $X_t, \Delta_t$ on the $t$-th factors $S, A$ describe the state and action at the $t$-th stage of the process($t \geqq 0$). Let $\Pi$ 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 $\gamma$. We denote by $F$ the set of functions on $S$ with $f(i) \in A$ for all $i \in S$. A randomized stationary policy $\gamma$ 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 $f$.

For any $X_0 = i, \pi \in \Pi$ and $q = (q_{ij}(a)) \in \mathbb{Q}$, we assume that $P(X_{t+1} = j|X_0, \Delta_0, \ldots, X_t = i, \Delta_t = a) = q_{ij}(a)$ 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 $\Omega$.

For a given reward function $r$ on $S \times A$, we shall consider the long-run expected average reward associated with $q \in \mathbb{Q}$ :

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

where $E_\pi(\cdot|X_0 = i, q)$ is the expectation operator w.r.t. $P_\pi(\cdot|X_0 = i, q)$. 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 by $\psi(i, q)$ the value function, i.e.,

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

$\pi^* \in \Pi$ 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 $\pi^*$ 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 $\lim_{n\to\infty} \psi(i, q|\pi^n) = \psi(i, q)$ for all $q \in \mathcal{D}$. Let $\mathbb{Q}^+ := \{q = (q_{ij}(a)) \in \mathbb{Q} \mid q_{ij}(a) > 0$ for all $i, j \in S$ and $a \in A\}$. In [10], a learning algorithm of the reward-penalty type (cf. [11]) was given, by which an adaptively optimal policy for $\mathbb{Q}^+$ was constructed by applying value iteration and policy improvement algorithms (cf. [3, 4, 5]). In this paper, we treat with the communicating case of multi-chain MDPs applying the idea of [10] extensively. The transition matrices $q = (q_{ij}(a)) \in \mathbb{Q}$ is said communicating (cf. [1, 6, 17]) if for any $i, j \in S$ there exists a path from $i$ to $j$ with positive probability, i.e., it holds that $q_{i_1i_2}(a_1)q_{i_2i_3}(a_2)\cdots q_{i_{l-1}i_l}(a_{l-1}) > 0$ 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 $\gamma$ defines an irreducible Markov chain(cf. [7]) 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. [17, 18]).

**Lemma 1.1** ([17, 18]). *Let $q = (q_{ij}(a)) \in \mathbb{Q}$. Supposed that there exists a constant $g$ and a $v \in B(S)$ such that*

(3) $$v(i) = \max_{a\in A}\{r(i, a) + \textstyle\sum_{j\in S} q_{ij}(a)v(j)\} - g \quad \text{for all } i \in S.$$

*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 (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

(4) $$v_\tau(i, q|\pi) = E_\pi(\textstyle\sum_{t=0}^{\infty}(1 - \tau)^t r(X_t, \Delta_t)|X_0 = i, q) \quad \text{for } i \in S, q \in \mathbb{Q} \text{ and } \pi \in \Pi,$$

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

(5) $$U_\tau\{q\}u(i) = \max_{a\in A}\{r(i, a) + (1 - \tau)\textstyle\sum_{j\in S} q_{ij}(a)u(j)\} \quad \text{for all } i \in S \text{ and } u \in B(S).$$

We have the following.

**Lemma 1.2** ([17, 18]). *It holds that (i) the operator $U_\tau\{q\}$ is a contraction with the modulus $(1-\tau)$ and (ii) the $(1-\tau)$-discount value function $v_\tau(i,q)$ is a unique fixed point of $U_\tau\{q\}$, i.e.,*

$$v_\tau = U_\tau\{q\}v_\tau, \tag{6}$$

*(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 $f_\tau$ is a maximizer of the right-hand side in (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*

$$\limsup_{\tau\to 0}|v_\tau(i,q) - v_\tau(j,q)| \leqq M \quad \text{for all} \quad i,j \in S. \tag{7}$$

*Proof.* See Appendix.■

Let $P(S)$ be the set of all probability distributions on $S$, i.e., $P(S) = \{\mu = (\mu_1, \ldots, \mu_N) | \mu_i \geqq 0, \sum_{i=1}^N \mu_i = 1$ for all $i \in S\}$. Let $q = (q_{ij}(a)) \in \mathbb{Q}$. For any $\tau \in (0,1)$ and $\mu = (\mu_1, \mu_2, \ldots, \mu_N) \in P(S)$, we perturb $q$ to $q^{\tau,\mu} = (q_{ij}^{\tau,\mu}(a))$ which is defined by

$$q_{ij}^{\tau,\mu}(a) = \tau\mu_j + (1-\tau)q_{ij}(a) \quad \text{for } i,j \in S \text{ and } a \in A. \tag{8}$$

The matrix expression of (8) is $q^{\tau,\mu} = \tau e\mu + (1-\tau)q$, where $e = (1,1,\ldots,1)^t$ is a transpose of $N$-dimensional vector $(1,1,\ldots,1)$. Then, we find that (6) in Lemma 1.2 can be rewritten as follows.

$$v_\tau(i,q) = \max_{a\in A}\{r(i,a) + \textstyle\sum_{j\in S} q_{ij}^{\mu,\tau}(a)v_\tau(j,q)\} - \tau\textstyle\sum_{j\in S}\mu_j v_\tau(j,q) \quad \text{for all } i \in S. \tag{9}$$

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) $f_\tau$ is $q^{\tau,\mu}$-optimal, where $f_\tau$ 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 $\tau$-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:*

$$u(i) = \max_{a\in A}\{r(i,a) + \textstyle\sum_{j\in S} q_{ij}(a)u(j)\} - \psi(q) \quad (i \in S), \tag{10}$$

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

*Proof.* See Appendix.■

We note that (i) in Theorem 2.1 derives the single average optimality equation for the communicating MDPs, which has been given first by [1]. In general, the value function $\psi(i,q)$ is known to be continuous on each equivalent class of $\mathbb{Q}$ (cf. [19, 20]), but (ii) in Theorem 2.1 gives as 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.

$$(11) \qquad \sigma^0(i,a) = 0, \sigma^n(i,a) = \inf\{t | t > \sigma^{n-1}(i,a), X_t = i, \Delta_t = a\} \ (n \geqq 1).$$

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 [9] to the communicating case.

**Lemma 3.1.** *Let $\pi = (\pi_0, \pi_1, \ldots)$ be any policy satisfying that there exists a decreasing sequence of positive numbers $\{\varepsilon_t\}_{t=0}^{\infty}$ such that (i) for each $t \geqq 0, \pi_t(a|h_t) \geqq \varepsilon_t$ for all $a \in A$ and $h_t = (x_0, a_0, \ldots, x_t) \in H_t$ and (ii) $\sum_{t=0}^{\infty} \varepsilon_t^N = \infty$. Then, $P_\pi(W | X_0 = i, q) = 1$ for all $q \in \mathbb{Q}^*$ and $i \in S$.*

*Proof.* See Appendix.∎

We note that as a example, the sequence $\{t^{-\frac{1}{N}}\}_{t=0}^{\infty}$ satisfies (ii) of Lemma 3.1.

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 $q_{ij}^n(a) = N_n(i,j|a)/N_n(i|a)$ if $N_n(i|a) > 0$, 0 otherwise. Then, $q_{ij}^n = \big(q_{ij}^n(a)\big)$ is the maximum likelihood estimator of the unknown transition matrices.

For any given $q^0 = \big(q_{ij}^0(a)\big) \in \mathbb{Q}$, we define $\tilde{q}^n = \big(\tilde{q}_{ij}^n(a)\big) \in \mathbb{Q}$ by $\tilde{q}_{ij}^n(a) = q_{ij}^n(a)$ if $N_n(i|a) > 0$, $q_{ij}^0(a)$ otherwise. We consider the following iterative scheme which is a variant of the non-stationary value iteration scheme proposed by [3]:

$$(12) \qquad \tilde{v}_0 = 0, \tilde{v}_{n+1} = U_\tau\{\tilde{q}^n\}\tilde{v}_n \quad (n \geqq 0).$$

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 (12). 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 $\phi$ 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 $\phi$. For each $n (n \geqq 0)$, letting $\tilde{\pi}_n^\tau(k|i) = P(\Delta_n = k | X_0, \Delta_0, \ldots, X_n = i)$ we propose to update $\tilde{\pi}_n^\tau$ as follows:
If, for each $i \in S, \tilde{a}_{n+1}(i) = a_i$,

$$(13) \qquad \tilde{\pi}_{n+1}^\tau(a_i|i) = 1 - \textstyle\sum_{\alpha \neq a_i} \phi(\tilde{\pi}_n^\tau(\alpha|i)), \tilde{\pi}_{n+1}^\tau(\alpha|i) = \phi(\tilde{\pi}_n^\tau(\alpha|i)) \ (\alpha \neq a_i).$$

In (13), 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 (13) is a learning algorithm of the reward-penalty type(cf. [11, 16, 21]). 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 (12) and (13).

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 $i \in S, a \in A$.

**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 (6).*

*Proof.* See Appendix.▮

Let ${}^{\tau}\mathbb{Q}^* := \{q^{\tau,\mu} | \mu \in P(S) \text{ and } q \in \mathbb{Q}^*\}$, where $q^{\tau,\mu}$ is defined in (8). Then, observing the discussion in Section 2 and ${}^{\tau}\mathbb{Q}^* \subset \mathbb{Q}^*$, from Lemma 3.2 we find that the results in [10] 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.* See Appendix.▮

## 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.

| $i$ | $a$ | $j=1$ | $j=2$ | $j=3$ | $r(i,a)$ |
|---|---|---|---|---|---|
| 1 | 1 | $p_1$ | $1-p_1$ | 0 | 3 |
| | 2 | $1-p_2$ | $p_2$ | 0 | 2.5 |
| 2 | 1 | 0 | $q_1$ | $1-q_1$ | 2 |
| | 2 | $1-q_2$ | $q_2$ | 0 | 1.5 |
| 3 | 1 | 0 | 0 | 1 | 1 |
| | 2 | 0 | 1 | 0 | 0.5 |

Table 4.1: parameterized transition matrices and reward function of simulated MDPs, where $i, j \in S$ and $a \in A$.

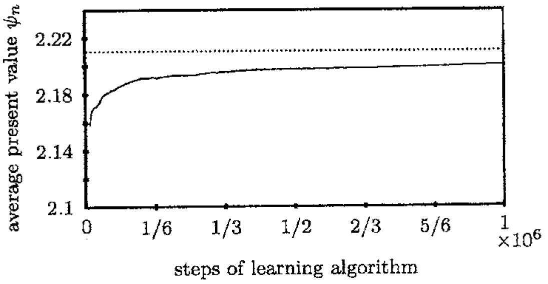

Figure 4.1: The trajectories of $\tilde{\psi}_n$ $(\tau = 0.01)$. The dotted line means the true optimal value of average reward.

We denote by $\tilde{\psi}_n$ the average present value until $n$-th time, which is defined by $\tilde{\psi}_n = \frac{1}{n}\sum_{t=0}^{n-1} r(X_t, \Delta_t)$ $(n \geqq 1)$.

We set $\tilde{\pi}_0^{\tau}(\cdot|i) = (\frac{1}{2}, \frac{1}{2})$ for each $i \in S$ and $q_0$ with $p_1 = \frac{2}{5}, q_1 = \frac{1}{2}, p_2 = \frac{3}{10}, q_2 = \frac{3}{10}$. We use a strictly increasing function $\phi$ such that $\phi(x) = (\frac{x^N}{1+x^N})^{\frac{1}{N}}$ where $N$ denotes the number of states in $S$. Note that it is easily checked that (i) in Condition A is satisfied by $\phi$ defined above when $b_0 = 1$ and $b_{n+1} = \phi(b_n)$ $(n \geqq 0)$.

Now, we make numerical experiments with the true transition matrices whose parameters are given by $p_1 = p_2 = \frac{1}{3}, q_1 = q_2 = \frac{2}{5}$. The corresponding trajectories of $\tilde{\psi}_n$ obtained by computer simulations of the learning algorithms (12)–(13) in Section 3 are given in Table 4.2 and Figure 4.1. It is shown that the optimal stationary policy of MDPs with true transition matrices such that $p_1 = p_2 = \frac{1}{3}, q_1 = q_2 = \frac{2}{5}$ are $\pi_t^*(1|1) = \pi_t^*(2|2) = \pi_t^*(2|3) = 1$ $(t \geqq 0)$ and the true optimal average reward $\frac{42}{19} \fallingdotseq 2.2105$. Also, adaptive decision rules and relative frequency of $N(i, j|a)$ at each $n$-step are listed in Table 4.3 and 4.4. The results of the above simulation show that the learning algorithm is practically effective for the communicating class of transition matrices.

| values | $n$<br>$\tau$ | $10^3$ | $5 \times 10^3$ | $10^4$ | $5 \times 10^4$ | $10^5$ | $10^6$ |
|---|---|---|---|---|---|---|---|
| $\tilde{\psi}_n$ | 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 |
| $\tilde{p}_1$ | 0.5 | 0.3156 | 0.3184 | 0.3189 | 0.3264 | 0.3292 | 0.3329 |
| | 0.2 | 0.3239 | 0.3198 | 0.3196 | 0.3266 | 0.3293 | 0.3329 |
| | 0.1 | 0.3221 | 0.3195 | 0.3195 | 0.3266 | 0.3293 | 0.3329 |
| | 0.01 | 0.3201 | 0.3195 | 0.3195 | 0.3266 | 0.3293 | 0.3329 |
| $\tilde{q}_1$ | 0.5 | 0.3714 | 0.3801 | 0.3914 | 0.3824 | 0.3833 | 0.3927 |
| | 0.2 | 0.3438 | 0.3738 | 0.3878 | 0.3811 | 0.3824 | 0.3926 |
| | 0.1 | 0.3438 | 0.3738 | 0.3878 | 0.3811 | 0.3824 | 0.3926 |
| | 0.01 | 0.3492 | 0.3756 | 0.3889 | 0.3815 | 0.3827 | 0.3926 |
| $\tilde{p}_2$ | 0.5 | 0.3372 | 0.3333 | 0.3234 | 0.3158 | 0.3343 | 0.3347 |
| | 0.2 | 0.3333 | 0.3305 | 0.3220 | 0.3152 | 0.3336 | 0.3345 |
| | 0.1 | 0.32 | 0.3247 | 0.3182 | 0.3137 | 0.3326 | 0.3344 |
| | 0.01 | 0.3165 | 0.3223 | 0.3190 | 0.3142 | 0.3324 | 0.3342 |
| $\tilde{q}_2$ | 0.5 | 0.3753 | 0.3931 | 0.3952 | 0.3969 | 0.3972 | 0.3992 |
| | 0.2 | 0.3695 | 0.3912 | 0.3943 | 0.3967 | 0.3971 | 0.3992 |
| | 0.1 | 0.3695 | 0.3912 | 0.3943 | 0.3967 | 0.3971 | 0.3992 |
| | 0.01 | 0.3741 | 0.3923 | 0.3948 | 0.3968 | 0.3972 | 0.3992 |

Table 4.2: The simulation values of $\tilde{\psi}_n$ for each discount parameter $\tau$ and step number $n$, and maximum likelihood estimates $\tilde{p}_1, \tilde{q}_1, \tilde{p}_2, \tilde{q}_2$ of $p_1, q_1, p_2, q_2$, where parameters are assumed such that $p_1 = p_2 = \frac{1}{3}, q_1 = q_2 = \frac{2}{5}$ and then the true optimal value of average reward is $\frac{42}{19} \fallingdotseq 2.2105$.

| decision rules | $n$<br>$\tau$ | $10^3$ | $5 \times 10^3$ | $10^4$ | $5 \times 10^4$ | $10^5$ | $10^6$ |
|---|---|---|---|---|---|---|---|
| $\tilde{\pi}_n^\tau(1\|1)$ | 0.5 | 0.9003 | 0.9416 | 0.9536 | 0.9729 | 0.9785 | 0.99 |
| | 0.2 | 0.8980 | 0.9413 | 0.9535 | 0.9728 | 0.9785 | 0.99 |
| | 0.1 | 0.8983 | 0.9413 | 0.9535 | 0.9728 | 0.9785 | 0.99 |
| | 0.01 | 0.8937 | 0.9409 | 0.9533 | 0.9728 | 0.9784 | 0.99 |
| $\tilde{\pi}_n^\tau(2\|2)$ | 0.5 | 0.8996 | 0.9415 | 0.9536 | 0.9729 | 0.9785 | 0.99 |
| | 0.2 | 0.9002 | 0.9415 | 0.9536 | 0.9729 | 0.9785 | 0.99 |
| | 0.1 | 0.9002 | 0.9415 | 0.9536 | 0.9729 | 0.9785 | 0.99 |
| | 0.01 | 0.9002 | 0.9415 | 0.9536 | 0.9729 | 0.9785 | 0.99 |
| $\tilde{\pi}_n^\tau(2\|3)$ | 0.5 | 0.9002 | 0.9415 | 0.9536 | 0.9729 | 0.9785 | 0.99 |
| | 0.2 | 0.9002 | 0.9415 | 0.9536 | 0.9729 | 0.9785 | 0.99 |
| | 0.1 | 0.9002 | 0.9415 | 0.9536 | 0.9729 | 0.9785 | 0.99 |
| | 0.01 | 0.9002 | 0.9415 | 0.9536 | 0.9729 | 0.9785 | 0.99 |

Table 4.3: Adaptive decision rules for each $n$-step

# Appendix

## Proof of Lemma 2.1

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 $\sigma^j$ by $\sigma^j = \sigma^j(H_t) =$ first $t \geqq 0$ 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 $f_\tau$ given in Lemma 1.2 the policy $\pi^j = (\pi_0^j, \pi_1^j, \ldots)$ will be defined by $\pi_t^j(\cdot|H_t) = \gamma(\cdot|X_t)$ if $t < \sigma^j(H_t)$, $f_\tau(X_t)$ if $t \geqq \sigma^j(H_t)$, $(t \geqq 0)$. Then we have the following:

$$(14) \quad v_\tau(i, q|\pi^j) = E_\gamma(\textstyle\sum_{t=0}^{\sigma^j - 1}(1-\tau)^t r(X_t, \Delta_t)|X_0 = i, q)$$
$$+ E_\gamma((1-\tau)^{\sigma^j}|X_0 = i, q) v_\tau(j|q) \quad (i \in S).$$

From irreducibility of the Markov chain induced by $q(\gamma)$, it holds (cf. [7]) that

$$(15) \qquad E_\gamma(\sigma^j|X_0 = i, q) < \infty \quad \text{for all} \quad i \in S.$$

Concerning with the second term of the right-hand side in (14), since $\lim_{\tau \to 0} \frac{(1-\tau)^n - 1}{\tau} = -n$ $(n \geqq 1)$, we have that

| $N_n(i,j\|a)/n$ | $\tau$ \ $n$ | $10^3$ | $5 \times 10^3$ | $10^4$ | $5 \times 10^4$ | $10^5$ | $10^6$ |
|---|---|---|---|---|---|---|---|
| $N_n(1,1\|1)/n$ | 0.5 | 0.1129 | 0.1290 | 0.1336 | 0.1446 | 0.1483 | 0.1540 |
| | 0.2 | 0.1149 | 0.1294 | 0.1338 | 0.1447 | 0.1483 | 0.1541 |
| | 0.1 | 0.1149 | 0.1294 | 0.1338 | 0.1447 | 0.1483 | 0.1541 |
| | 0.01 | 0.1049 | 0.1274 | 0.1328 | 0.1445 | 0.1482 | 0.1540 |
| $N_n(1,2\|1)/n$ | 0.5 | 0.2448 | 0.2761 | 0.2853 | 0.2985 | 0.3020 | 0.3087 |
| | 0.2 | 0.2398 | 0.2751 | 0.2848 | 0.2983 | 0.3020 | 0.3087 |
| | 0.1 | 0.2418 | 0.2755 | 0.2850 | 0.2984 | 0.3020 | 0.3087 |
| | 0.01 | 0.2228 | 0.2713 | 0.2828 | 0.2979 | 0.3018 | 0.3087 |
| $N_n(1,1\|2)/n$ | 0.5 | 0.0569 | 0.0284 | 0.0226 | 0.0125 | 0.0095 | 0.0046 |
| | 0.2 | 0.0679 | 0.0312 | 0.0240 | 0.0128 | 0.0097 | 0.0046 |
| | 0.1 | 0.0679 | 0.0312 | 0.0240 | 0.0128 | 0.0097 | 0.0046 |
| | 0.01 | 0.0949 | 0.0370 | 0.0269 | 0.0134 | 0.0100 | 0.0046 |
| $N_n(1,2\|2)/n$ | 0.5 | 0.0290 | 0.0142 | 0.0108 | 0.0058 | 0.0048 | 0.0023 |
| | 0.2 | 0.0340 | 0.0154 | 0.0114 | 0.0059 | 0.0048 | 0.0023 |
| | 0.1 | 0.0320 | 0.0150 | 0.0112 | 0.0059 | 0.0048 | 0.0023 |
| | 0.01 | 0.0440 | 0.0176 | 0.0126 | 0.0061 | 0.0050 | 0.0023 |
| $N_n(2,2\|1)/n$ | 0.5 | 0.0260 | 0.0168 | 0.0137 | 0.0079 | 0.0061 | 0.0031 |
| | 0.2 | 0.0220 | 0.0160 | 0.0133 | 0.0078 | 0.0061 | 0.0031 |
| | 0.1 | 0.0220 | 0.0160 | 0.0133 | 0.0078 | 0.0061 | 0.0031 |
| | 0.01 | 0.0220 | 0.0160 | 0.0133 | 0.0078 | 0.0061 | 0.0031 |
| $N_n(2,3\|1)/n$ | 0.5 | 0.0440 | 0.0274 | 0.0213 | 0.0128 | 0.0099 | 0.0047 |
| | 0.2 | 0.0420 | 0.0268 | 0.0210 | 0.0127 | 0.0099 | 0.0047 |
| | 0.1 | 0.0420 | 0.0268 | 0.0210 | 0.0127 | 0.0099 | 0.0047 |
| | 0.01 | 0.0410 | 0.0266 | 0.0209 | 0.0127 | 0.0098 | 0.0047 |
| $N_n(2,1\|2)/n$ | 0.5 | 0.2727 | 0.2901 | 0.2961 | 0.3042 | 0.3068 | 0.3110 |
| | 0.2 | 0.2727 | 0.2903 | 0.2962 | 0.3042 | 0.3068 | 0.3110 |
| | 0.1 | 0.2727 | 0.2903 | 0.2962 | 0.3042 | 0.3068 | 0.3110 |
| | 0.01 | 0.2657 | 0.2887 | 0.2954 | 0.3041 | 0.3067 | 0.3110 |
| $N_n(2,2\|2)/n$ | 0.5 | 0.1638 | 0.1880 | 0.1935 | 0.2002 | 0.2022 | 0.2066 |
| | 0.2 | 0.1598 | 0.1866 | 0.1928 | 0.2001 | 0.2021 | 0.2066 |
| | 0.1 | 0.1598 | 0.1866 | 0.1928 | 0.2001 | 0.2021 | 0.2066 |
| | 0.01 | 0.1588 | 0.1864 | 0.1927 | 0.2001 | 0.2021 | 0.2066 |
| $N_n(3,3\|1)/n$ | 0.5 | 0.0060 | 0.0026 | 0.0019 | 0.0007 | 0.0005 | 0.0001 |
| | 0.2 | 0.0050 | 0.0024 | 0.0018 | 0.0007 | 0.0005 | 0.0001 |
| | 0.1 | 0.0050 | 0.0024 | 0.0018 | 0.0007 | 0.0005 | 0.0001 |
| | 0.01 | 0.0050 | 0.0024 | 0.0018 | 0.0007 | 0.0005 | 0.0001 |
| $N_n(3,2\|2)/n$ | 0.5 | 0.0440 | 0.0274 | 0.0213 | 0.0128 | 0.0099 | 0.0047 |
| | 0.2 | 0.0420 | 0.0268 | 0.0210 | 0.0127 | 0.0099 | 0.0047 |
| | 0.1 | 0.0420 | 0.0268 | 0.0210 | 0.0127 | 0.0099 | 0.0047 |
| | 0.01 | 0.0410 | 0.0266 | 0.0209 | 0.0127 | 0.0098 | 0.0047 |

Table 4.4: Relative frequency of $N(i,j|a)/n$

$$\begin{aligned} & \liminf_{\tau\to 0} \tfrac{1}{\tau}\{E_\gamma((1-\tau)^{\sigma^j}|X_0 = i,q) - 1\} \\ (16)\qquad & \geqq \textstyle\sum_{n=0}^{\infty} \liminf_{\tau\to 0} \frac{(1-\tau)^n - 1}{\tau} P_\gamma(\sigma^j = n \mid X_0 = i,q) \\ & = -\textstyle\sum_{n=1}^{\infty} n P_\gamma(\sigma^j = n \mid X_0 = i,q) = -E_\gamma(\sigma^j \mid X_0 = i,q). \end{aligned}$$

On the other hand, from (14) it holds that $v_\tau(i,q) - v_\tau(j,q) \geqq v_\tau(i,q|\pi^j) - v_\tau(j,q) \geqq -\|r\|E_\gamma(\sigma^j|X_0 = i,q) + \{E_\gamma((1-\tau)^{\sigma^j}|X_0 = i,q) - 1\}v_\tau(j,q)$ where $\|r\| = \max_{i\in S, a\in A}|r(i,a)|$. Thus, by (14), (16) and Lemma 1.2(iii) we have that $\liminf_{\tau\to 0}(v_\tau(i,q) - v_\tau(j,q)) \geqq -\limsup_{\tau\to 0}(\|r\| + |\tau v_\tau(j,q)|)E_\gamma(\sigma^j|X_0 = i,q) = -(\|r\| + |\psi(j,q)|)E_\gamma(\sigma^j|X_0 = i,q) > -\infty$. Similarly, we get that $\liminf_{\tau\to 0}(v_\tau(j,q) - v_\tau(i,q)) \geqq -(\|r\| + |\psi(i,q)|)E_\gamma(\sigma^i|X_0 = j,q) > -\infty$, and hence $\limsup_{\tau\to 0}(v_\tau(i,q) - v_\tau(j,q)) \leqq (\|r\| + |\psi(i,q)|)E_\gamma(\sigma^i|X_0 = j,q) < \infty$. If we put $M := \max_{i,j\in S}(\|r\| + |\psi(j,q)|)E_\gamma(\sigma^j|X_0 = i,q)$, (7) follows, which completes the proof.∎

**Proof of Theorem 2.1**

For any fixed $i_0 \in S$, let $u_\tau(j) = v_\tau(j,q) - v_\tau(i_0,q)$ for each $j \in S$. Then, from (9) we get

$$(17)\qquad u_\tau(i) = \max_{a\in A}\{r(i,a) + \textstyle\sum_{j\in S} q_{ij}^{\mu,\tau}(a)u_\tau(j)\} - \tau\sum_{j\in S}\mu_j v_\tau(j,q) \quad (i \in S).$$

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 (17) with $\tau = \tau_l$, we get (10) 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.∎

**Proof of Lemma 3.1**

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

$$(18) \quad P(X_t = i, \Delta_t = a \text{ for some } t \text{ with } n \leqq t \leqq n + N | H_n) \geqq \delta \varepsilon^N_{n+N} \quad \text{for any } n \geqq 0 \text{ and } i \in S, a \in A.$$

Let $B_t := \{\sigma^n(i, a) = t \text{ for some } n \geqq 1\}$. Then, we observe that $W(i, a) = \limsup_{t\to\infty} B_t = (\liminf_{t\to\infty} B^c_t)^c$, so that it holds that

$$(19) \qquad P(W(i, a)) = 1 - P(\liminf_{t\to\infty} B^c_t).$$

For any positive integer with $L > n$, let $l := [\frac{(L-n)}{N}]$, where for a real number $z$, $[z]$ is the largest integer equal to or less than $z$. Then, we have from (19) and (ii) in Lemma 3.1 that $P(\bigcap_{t=n}^{L} B^c_t) \leqq P(\bigcap_{\alpha=0}^{l} \bigcap_{t=n+\alpha N}^{n+(\alpha+1)N-1} B^c_t) \leqq \{1 - P(\bigcup_{t=n}^{n+N-1} B_t)\} \cdots \{1 - P(\bigcup_{t=lN}^{n+(l+1)N-1} B_t | \bigcap_{t=n}^{n+lN} B^c_t)\} \leqq (1 - \delta\varepsilon^N_{n+N-1}) \cdots (1 - \delta\varepsilon^N_{n+(l+1)N-1}) \leqq e^{-\delta \sum_{i=1}^{l+1} \varepsilon^N_{n+iN-1}} \to 0$ as $L \to \infty$, which implies that $\lim_{L\to\infty} P(\bigcap_{t=n}^{L} B^c_t) = P(\bigcap_{t=n}^{\infty} B^c_t) = 0$ for all $n \geqq 1$. Thus, from (19) $P(W(i, a)) = 1$, which implies $P(W) = 1$.∎

**Proof of Lemma3.2**

For each $i \in S$ and $a \in A$, we show by induction that $\tilde{\pi}^\tau_n(a|i) \geqq b_{n+m}$ $(n \geqq 1)$ for some $m \geqq 1$. By (i) of Condition A, there exists an integer $m$ for which $\tilde{\pi}^\tau_0(a|i) > b_m$ for all $i \in S$ and $a \in A$, such that it holds from (13) and the property of $\phi$ that $\tilde{\pi}^\tau_1(a|i) = \phi(\tilde{\pi}^\tau_0(a|i)) > \phi(b_m) = b_{m+1}$ for $a \neq \tilde{a}_1(i)$ and $\tilde{\pi}^\tau_1(\tilde{a}_1(i)|i) > \tilde{\pi}^\tau_0(\tilde{a}_1(i)|i) > b_m > b_{m+1}$. For $n > 1$, it holds from the assumption of the induction that $\tilde{\pi}^\tau_{n+1}(a|i) \geqq \phi(\tilde{\pi}^\tau_n(a|i)) \geqq \phi(b_{n+m}) = b_{n+m+1}$. Thus, from (i) of Condition A, Lemma 3.1 shows that $P_{\tilde{\pi}^\tau}(W|X_0 = i, q) = 1$ for $q \in \mathbb{Q}^*$, which means that $\lim_{n\to\infty} N_n(i|a) = \infty$ $P_{\tilde{\pi}^\tau}(\cdot|X_0 = i, q)$-a.s. By applying the law of large numbers(cf. [2]), we get (i). Also, (ii) and (iii) follows clearly from (i).∎

**Proof of Theorem 3.2**

Let $q \in \mathbb{Q}^*$. For each $t \geqq 0$, let $\tilde{\delta}_t := (1-\tau)\tilde{v}_t(X_t) - \{r(X_t, \Delta_t) + (1-\tau)\tilde{v}_t(X_{t+1})\}$ and$\delta_t(j) := E_{\tilde{\pi}^\tau}(\tilde{\delta}_t|H_t, X_t = j, q)$. Then, by the stability theorem(cf. [13]), we get

$$(20) \qquad \lim_{T\to\infty} \tfrac{1}{T+1} \textstyle\sum_{t=0}^{T}\{\tilde{\delta}_t - \delta_t(X_t)\} = 0, \;\; P_{\tilde{\pi}^\tau}(\cdot|X_0 = i, q)\text{-a.s.}$$

On the other hand, it holds that $\delta_t(j) = \tilde{v}_t(j) - \sum_{a\in A}\{\{r(j, a) + (1-\tau)\sum_{k\in S} q_{jk}(a)\tilde{v}_t(k)\}$ $\tilde{\pi}^\tau_t(a|j)\} - \tau\tilde{v}_t(j)$. So, by (ii) and (iii) of Lemma 3.2, $\lim_{t\to\infty} \delta_t(j) = -\tau v_\tau(j, q)$, $P_{\tilde{\pi}^\tau}(\cdot|X_0 = i, q)$-a.s. Thus, from (20) it holds that

$$(21) \qquad \begin{aligned} \min_{i\in S} -\tau v_\tau(i, q) &\leqq \liminf_{T\to\infty} \tfrac{1}{T+1} \textstyle\sum_{t=0}^{T} \tilde{\delta}_t \\ &\leqq \limsup_{T\to\infty} \tfrac{1}{T+1} \textstyle\sum_{t=0}^{T} \tilde{\delta}_t \leqq \max_{i\in S} -\tau v_\tau(i, q). \end{aligned}$$

However, $\sum_{t=0}^{T} \tilde{\delta}_t = -\sum_{t=0}^{T} r(X_t, \Delta_t) + (1-\tau)\sum_{t=0}^{T}(\tilde{v}_t(X_t) - \tilde{v}_t(X_{t+1}))$, so that by (ii) of Lemma 3.2 $\limsup_{T\to\infty}(\liminf_{T\to\infty}) \frac{1}{T+1}\sum_{t=0}^{T}\tilde{\delta}_t = \limsup_{T\to\infty}(\liminf_{T\to\infty}) - \frac{1}{T+1}\sum_{t=0}^{T} r(X_t, \Delta_t)$. Thus, applying Fatou's Lemma, from (21) we get

$$(22) \qquad \min_{i\in S} \tau v_\tau(i, q) \leqq \psi(i, q|\tilde{\pi}^\tau) \leqq \max_{i\in S} \tau v_\tau(i, q).$$

By Lemma 1.2 and Theorem 2.1 (i), $\lim_{\tau\to 0} \tau v_\tau(i, q) = \psi(q)$, which implies from (22) that $\psi(i, q|\tilde{\pi}^\tau) \to \psi(q)$ as $\tau \to 0$. This completes the proof.∎

## References

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

[2] Patrick Billingsley. *Statistical inference for Markov processes.* Statistical Research Monographs, Vol. II. The University of Chicago Press, Chicago, Ill., 1961.

[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. *(preprint)*, 2005.

[7] John G. Kemeny and J. Laurie Snell. *Finite Markov chains.* The University Series in Undergraduate Mathematics. D. Van Nostrand Co., Inc., Princeton, N.J.-Toronto-London-New York, 1960.

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

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

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

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

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

[13] Michel Loève. *Probability theory.* Third edition. D. Van Nostrand Co., Inc., Princeton, N.J.-Toronto, Ont.-London, 1963.

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

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

[16] M. R. Meybodi and S. Lakshmivarahan. $\varepsilon$-optimality of a general class of learning algorithms. *Inform. Sci.*, 28(1):1–20, 1982.

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

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

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

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

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

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