Temporal Difference-based Adaptive policies in Neuro-dynamic Programming

T. IKI[1], M. HORIGUCHI[2], M. YASUDA[3] and M. KURANO[4]

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

Abstract. Based on temporal difference method in neuro-dynamic programming, an adaptive policy for finite state Markov decision processes with the average reward is constructed under the minorization condition. We estimate the value function by a learning iteration algorithm. And the adaptive policy is specified as an ε-forced modification of the greedy policy for the estimated value and the estimated transition probability matrix. Also, a numerical experiment for "Toymaker's problem" is given to illustrate the validity of the adaptive policy.

Keywords: adaptive Markov decision processes, temporal difference method, average reward, minorization condition, neuro-dynamic programming.

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 [6, 12, 14, 15] and so on. The idea of neuro-dynamic programming by [2] seems a recent breakthrough in the practical application of neural networks and dynamic programming to complex problems of planning, optimal decision making, and intelligent control. In our previous work [9], the adaptive policies are constructed by applying the methods of value iteration, cooperated with the policy improvement(cf. [17]), in which the corresponding value function is approximated through the expectation with respect to the estimated transition matrices at each learning step. In order to decrease the amount of computation necessary for a learning algorithm, we can use the idea of temporal-difference method(TD-method) in neuro-dynamic programming[1, 2, 4, 10, 18], that is, the data of the state-action process are directly used for value iteration and policy improvement at each learning step, by which the amount of computation is reduced to some extent.

In this paper, based on TD-method, we propose a way of constructing an adaptive policy for the average reward criterion. In fact, we will estimate the value function by the stochastic iteration algorithm that updates the value estimate iteratively using temporal differences made from the data of state-action process and adaptively optimal policies for a class of MDPs satisfying the minorization condition are constructed as an ε-forced modification of the greedy policy for the estimated value and transition probability matrices. Also, a numerical experiment for "Toymaker's problem" in Howard[7] is given to illustrate the validity of the adaptive policy we have proposed here.

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.

For any $\delta > 0$, let $\mathbb{Q}_\delta$ denote the parameter space of K unknown stochastic matrices, defined by

$$\mathbb{Q}_\delta = \left\{ q = (q_{ij}(a)) \;\middle|\; q_{ij}(a) \geqq \delta, \sum_{j \in S} q_{ij}(a) = 1 \;\text{ for }\; i, j \in S, a \in A \right\}.$$

Throughout this paper, all the tranisition matrices which we are dealing with are satisfying the minorization condition above(cf. [16]).

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 $\xi = (\xi(\cdot|i) : i \in S) \in P(A|S)$ such that $\pi_t(\cdot|x_0, a_0, \ldots, x_t) = \xi(\cdot|x_t)$ 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 ξ. We denote by F the set of functions on S with $f(i) \in A$ for all $i \in S$. A randomized stationary policy ξ is called stationary if there exists a function $f \in F$ with $\xi(\{f(i)\}|i) = 1$ 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:

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

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

Then, for any fixed $\delta > 0$, the problem is to maximize $\psi(i, q|\pi)$ over all $\pi \in \Pi$ for $i \in S$ and $q \in \mathbb{Q}_\delta$.

Thus, for $q \in \mathbb{Q}_\delta$ denoting by $\psi(i,q)$ the value function, i.e.,

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

$\pi^* \in \mathit{\Pi}$ will be called q-optimal if $\psi(i,q|\pi^*) = \psi(i,q)$ for all $i \in S$ and called adaptively optimal if π^* is q-optimal for all $q \in \mathbb{Q}_\delta$.

In Section 2, several lemmas are prepared, which are used to prove the validity of adaptive policies. These notions are proposed in Section 3. 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 $B(S)$ be the set of all functions on S. The following fact is well-known (cf. [17]).

Lemma 1 (Puterman [17]). *Let $q = (q_{ij}(a)) \in \mathbb{Q}_\delta$. Supposed that there exists a constant g and a $v \in B(S)$ such that*

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

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 (3) for all $i \in S$.

For any $q \in \mathbb{Q}_\delta$, we define the map $U\{q\} : B(S) \to B(S)$ by

$$U\{q\}u(i) = \max_{a \in A}\left\{ r(i,a) + \sum_{j \in S}\big(q_{ij}(a) - \delta\big)u(j) \right\} \tag{4}$$

Then, $U(q)$ is easily proved to be a contraction mapping, so that there is a fixed point. Let $h(q) \in B(S)$ be unique fixed point of $U(q)$. We have the optimality equation for the average case:

$$h(q) = U\{q\}h(q) \quad \text{for each} \quad q \in \mathbb{Q}_\delta. \tag{5}$$

Putting $\psi^*(q) = \delta \sum_{j \in S} h(q)(j)$ in (5), we obtain the optimality equation:

$$h(q)(i) = \max_{a \in A}\left\{ r(i,a) - \psi^*(q) + \sum_{j \in S} q_{ij}(a)h(q)(j) \right\} \tag{6}$$

for $i \in S$. Then, by applying Lemma 1, we have the well-known results.

Lemma 2. *Let $q \in \mathbb{Q}_\delta$. Then, $\psi^*(q) = \psi(i,q)$ (independent of $i \in S$) and if $f(i) \in A^*(i|q)$ for all $i \in S$, f is q-optimal, where $A^*(i|q)$ is the set of optimal actions at state i and $A^*(i|q) = \arg\max_{a \in A}\left\{ r(i,a) - \psi^*(q) + \sum_{j \in S} q_{ij}(a)h(q)(j) \right\}$*

For any map $H : B(S) \to B(S)$, we consider the stochastic algorithm $\{\tilde{v}_t : t = 0, 1, 2, \cdots\}$ for the stochastic process $\{X_t\}_{t=0}^{\infty}$ on S, whose update equations are described by, for $i \in S$,

$$\begin{aligned} &\tilde{v}_0(i) \equiv 0, \\ &\tilde{v}_{t+1}(i) = \big(1 - \tilde{\gamma}_t(i)\big)\tilde{v}_t(i) + \tilde{\gamma}_t(i)\Big(H\tilde{v}_t(i) + W_t(i) + u_t(i)\Big), \quad t \geqq 0 \end{aligned} \tag{7}$$

where $\tilde{\gamma}_t(i)$ is a step size at time t and defined for a given sequence $\{\gamma_t(i)\}$ by $\tilde{\gamma}_t(i) = \gamma_t(i)$ if $X_t = i$ and $= 0$ otherwise. Also, $\{W_t(i)\}$ and $\{u_t(i)\}$ are random noise terms depending on $i \in S$.

Then, we have the following.

Lemma 3 (cf. Proposition 4.5 in [2]). *Suppose that the following condition (i) – (v).*

(i) $E[W_t(i)|\mathcal{F}_t] = 0$ *for* $i \in S$

(ii) There exist $A, B > 0$ *such that*

$$E\big[W_t(i)^2 \;\big|\; \mathcal{F}_t\big] \leqq A + B\|\tilde{v}_t\|^2$$

for $t \geqq 0$ *and* $i \in S$.

(iii) H *is a contraction with a unique fixed point* $v^* \in B(S)$.

(iv) $\tilde{\gamma}_t(i) \geqq 0$, $\sum_{t=0}^{\infty} \tilde{\gamma}_t(i) = \infty$ *and* $\sum_{t=0}^{\infty} \tilde{\gamma}_t(i)^2 < \infty$ *for* $t \geqq 0$, $i \in S$.

(v) There exists a nonnegative random sequence θ_t *that converges to zero with probability 1, and such that*

$$|u_t(i)| \leqq \theta_t(||\tilde{v}_t|| + 1)$$

for $i \in S$ *and* $t \geqq 0$.

Then, $\tilde{v}_t$ *in (7) converges to* v^* *with probability 1, where* $||\cdot||$ *is a supremum norm and* $\mathcal{F}_t$ *is a minimal* σ*-field generated by* $\Big\{X_\ell(\ell \leqq t), W_\ell(\ell \leqq t-1), U_\ell(\ell \leqq t-1)\Big\}$.

3 TD-based adaptive policies

In this section, an adaptive policy is given by the learning iteration algorithm using temporal differences.

Lemma 4. *Let* $\pi = (\pi_0, \pi_1, \cdots) \in \Pi$ *satisfy that* $\pi_t\big(A^*(j|q) \;\big|\; X_0, \Delta_0, \cdots, \Delta_{t-1}, X_t = j\big) \to 1$ *with* $P_\pi(\cdot|X_0 = i, q)$*-probability* 1 *as* $t \to \infty$ *for* $i, j \in S$ *and* $q \in \mathbb{Q}_\delta$. *Then, the policy* π *is adaptively optimal for* $\mathbb{Q}_\delta$.

Proof. Let $q \in \mathbb{Q}_\delta$ with $\delta > 0$. For simplicity, put $P(\cdot) = P_\pi(\cdot|X_0 = i, q)$ and $E(\cdot) = E_\pi(\cdot|X_0 = i, q)$. For $t \geqq 0$, let

$$\sigma(X_t, X_{t+1}) = h(q)(X_t) - \left(r(X_t, \Delta_t) + h(q)(X_{t+1}) - \psi^*(q)\right).$$

Then, we have

$$E(\sigma(X_t, X_{t+1})|X_0, \Delta_0, \ldots, X_t = j)$$
$$= h(q)(j) - \sum_{a\in A}\left(r(j,a) + \sum_{\ell\in S} q_{j\ell}(a)h(q)(\ell) - \psi^*(q)\right)\pi_t(a|X_0 = i, \ldots, X_t = j). \tag{8}$$

By assumption of Lemma 4, $\pi_t(A^*(j|q)|X_0, \Delta_0, \ldots, X_t = j) \to 1$ $(t \to \infty)$ with P-probability 1. So, by (8) it holds that $E\left(\sigma(X_t, X_{t+1})|X_0, \Delta_0, \ldots, X_t = j\right) \to 0$ as $t \to \infty$ with P-probability 1. Applying the dominated convergence theorem, it follows that $E\left(\sigma(X_t, X_{t+1})\right) \to 0$ as $t \to \infty$. Thus, we have

$$\lim_{T\to\infty} \frac{1}{T+1}\sum_{t=0}^{T} E(\sigma(X_t, X_{t+1}))$$
$$= \lim_{T\to\infty}\left(\frac{1}{T+1}E\left(\sum_{t=0}^{T} r(X_t, \Delta_t)\right) - \psi^*(q) + \frac{1}{T+1}\left(h(q)(X_0) - h(q)(X_T)\right)\right)$$
$$= \psi(i, q|\pi) - \psi^*(q) = 0.$$

By Lemma 2, $\psi^*(q) = \psi(i, q) = \psi(i, q|\pi)$ which completes the proof. □

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

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

Firstly, the adaptive policy is constructed in the following TD-based learning algorithm "Algorithm $(*)$" with the sequences $\{\varepsilon_t(i)\}_{t=0}^{\infty}$ for each $i \in S$ such that $0 < \varepsilon_t(i) < 1$ for $t \geqq 0$ and $i \in S$.

Algorithm $(*)$:

1. Set $t = 0$ and $\tilde{v}_0 \equiv 0$ and let $\tilde{\pi}_0 \in P(A|S)$ be such that $\tilde{\pi}_0(a|i) > 0$ for all $a \in A$ and $i \in S$.
2. Suppose that $\tilde{\pi}_t \in P\big(A \mid (S \times A)^t \times S\big)$ and $\tilde{v}_t \in B(S)$ are given and Δ_t is chosen according to $\tilde{\pi}_t$. Observe the next state $X_{t+1} = j$ selected according to the present state $X_t = i$ and Δ_t.
   At the stage $t+1$, determine $\tilde{v}_{t+1} \in B(S)$ by the TD-based update equation: for $i \in S$,

$$\tilde{v}_{t+1}(i) = \big(1 - \tilde{\gamma}_t(i)\big)\tilde{v}_t(i) + \tilde{\gamma}_t(i)\Big(r(i, \Delta_t) + \tilde{v}_t(X_{t+1}) - \delta\sum_{\ell\in S}\tilde{v}_t(\ell)\Big) \tag{9}$$

where the step size $\tilde{\gamma}_t$ is defined as $\tilde{\gamma}_t(i) = \gamma_t(i)$ if $X_t = i$, and $= 0$ otherwise, for a sequence $\{\gamma_t(i)\}$.

3. Let $\tilde{a}_{t+1}(i) \in \arg\max_{a\in A}\left\{r(i,a)+\sum_{j\in S}\tilde{q}_{ij}^t(a)\tilde{v}_{t+1}(j)\right\}$ for each $i \in S$. Then the policy $\tilde{\pi}_{t+1}$ is given by

$$\tilde{\pi}_{t+1}(a|i) = \frac{\varepsilon_t(i)}{K(i)-1} \text{ if } a \neq \tilde{a}_{t+1}(i), = 1-\varepsilon_t(i) \text{ if } a = \tilde{a}_{t+1}(i), \tag{10}$$

where $K(i)$ denotes the number of actions in state i.

4. Set $t = t+1$ and return to step 2.

To prove the validity of the adaptive policy given in Algorithm $(*)$, we need the following Condition $(*)$ and Lemmas.

Condition $(*)$

(i) $\lim_{t\to\infty}\varepsilon_t(i) = 0$ and $\sum_{t=0}^{\infty}\varepsilon_t(i) = \infty$,
(ii) $\gamma_t(i) \geqq 0, \sum_{t=0}^{\infty}\gamma_t(i) = \infty$ and $\sum_{t=0}^{\infty}\gamma_t(i)^2 < \infty$ for all $i \in S$.

Lemma 5. *Suppose that (i) of Condition $(*)$ holds and $q = (q_{ij}(a)) \in \mathbb{Q}_\delta$ with $\delta > 0$. Then, we have*

(i) $\lim_{t\to\infty} N_t(j|a) = \infty$ *for all $j \in S$ and $a \in A$ with $P_{\tilde{\pi}}(\cdot|X_0 = i, q)$-probability 1,*
(ii) $q_{ij}^t(a) \to q_{ij}(a)$ *as $t \to \infty$ for all $i, j \in S$ and $a \in A$ with $P_{\tilde{\pi}}(\cdot|X_0 = i, q)$-probability 1.*

Proof. From Lemma 1 of [11], (i) follows. Also, (ii) follows from the law of large numbers(cf. [3]). □

Applying Lemma 3, we have the following.

Theorem 1. *Suppose that Condition $(*)$ holds and $q = (q_{ij}(a)) \in \mathbb{Q}_\delta$ with $\delta > 0$. Then, $\tilde{v}_t(i) \to h(q)(i)$ as $t \to \infty$ with $P_{\tilde{\pi}}(\cdot|X_0 = i, q)$-probability 1.*

Proof. For simplicity, put $P(\cdot) = P_{\tilde{\pi}}(\cdot|X_0 = i, q)$. We rewrite the update equation (9) in Algorithm $(*)$ to:

$$\tilde{v}_{t+1}(i) = (1-\tilde{\gamma}_t(i))\tilde{v}_t(i) + \tilde{\gamma}_t(i)\left(U\{q\}\tilde{v}_t(i) + W_t(i) + u_t(i)\right) \ (t \geqq 0), \tag{11}$$

where

$$W_t(i) = r(i, \Delta_t) + \tilde{v}_t(X_{t+1}) - \delta\sum_{\ell\in S}\tilde{v}_t(\ell) - \sum_{a\in A}\left(r(i,a) + \sum_{j\in S}q_{ij}(a)\tilde{v}_t(j) - \delta\sum_{\ell\in S}\tilde{v}_t(\ell)\right)\tilde{\pi}_t(a|i), \tag{12}$$

$$u_t(i) = \sum_{a\in A}\left(r(i,a) + \sum_{j\in S}q_{ij}(a)\tilde{v}_t(j) - \delta\sum_{\ell\in S}\tilde{v}_t(\ell)\right)\tilde{\pi}_t(a|i) - U\{q\}\tilde{v}_t(i) \ (t \geqq 0). \tag{13}$$

It is easily seen that $W_0(i) = u_0(i) = 0$ for $i \in S$.

It follows from the definition of $U\{q\}$ that $||U\{q\}\tilde{v}_t|| \leqq ||r|| + (1-\delta)||\tilde{v}_t||$ with $1-\delta < 1$, so that by Proposition 4.7 of [2], the sequence $\tilde{v}_t$ is bounded with P-probability 1. That is, there exists $M > 0$ such that $||\tilde{v}_t|| \leqq M$ for $t \geqq 1$ with P-probability 1.

By (12), $|W_t(i)| \leqq 2(||r|| + M)$ for all $t \geqq 1$ and $i \in S$, such that conditions (i)–(ii) of Lemma 3.

For u_t in (13), we have

$$|u_t(i)| \leqq A_1 + A_2 + A_3 \tag{14}$$

where

$$A_1 = \left| \sum_{a \in A} \left(r(i,a) + \sum_{j \in S} \tilde{q}_{ij}^{t-1}(a)\tilde{v}_t(j) - \delta \sum_{\ell \in S} \tilde{v}_t(\ell) \right) \tilde{\pi}_t(a|i) - U\{\tilde{q}^{t-1}\}\tilde{v}_t(i) \right|,$$

$$A_2 = \left| \sum_{a \in A} \left(r(i,a) + \sum_{j \in S} \tilde{q}_{ij}^{t-1}(a)\tilde{v}_t(j) \right) \tilde{\pi}_t(a|i) - \sum_{a \in A} \left(r(i,a) + \sum_{j \in S} q_{ij}(a)\tilde{v}_t(j) \right) \tilde{\pi}_t(a|i) \right|$$

and

$$A_3 = \left| U\{\tilde{q}^t\}\tilde{v}_t(i) - U\{q\}\tilde{v}_t(i) \right| \quad (t \geqq 1).$$

By the definition of $\tilde{\pi}_t$, $A_1 \leqq (||r|| + M)\,\varepsilon_t$ where $\varepsilon_t = \max_{i \in S}\{\varepsilon(i)\}$. We also observe that $A_2 \leqq M \max_{i \in S, a \in A} \sum_{j \in S} \left| \tilde{q}_{ij}^t(a) - q_{ij}(a) \right|$. Moreover, we have

$$\begin{aligned} A_3 &\leqq \max_{i \in S, a \in A} \left| \left(r(i,a) + \sum_{j \in S} \tilde{q}_{ij}^{t-1}\tilde{v}_t(j) \right) - \left(r(i,a) + \sum_{j \in S} q_{ij}(a)\tilde{v}_t(j) \right) \right| \\ &= M \max_{i \in S, a \in A} \sum_{j \in S} \left| \tilde{q}_{ij}^{t-1}(a) - q_{ij}(a) \right|. \end{aligned}$$

Thus, putting $\theta_t = \varepsilon_t + \max_{i \in S, a \in A} \sum_{j \in S} \left| \tilde{q}_{ij}^{t-1}(a) - q_{ij}(a) \right|$ and $C = 3M + ||r||$, it yields that $|u_t(i)| \leqq C\theta_t$ for all $t \geqq 0$.

By Lemma 5 and (i) of Condition $(*)$, it clearly holds that $\theta_t \to 0$ as $t \to \infty$ with P-probability 1, which implies condition (v) of Lemma 3. Thus, since $h(q)$ is a fixed point of $U\{q\}$, by applying Lemma 3, the desired results follows, which completes the proof. □

The optimality of the adaptive policy $\tilde{\pi}$ is given in the following.

Theorem 2. *Let $\delta > 0$ be arbitrary. Suppose that Condition $(*)$ holds. Then, $\tilde{\pi}$ is adaptively optimal for $\mathbb{Q}_\delta$.*

Proof. Let $q = (q_{ij}(a)) \in \mathbb{Q}_\delta$. By Theorem 1, $\tilde{v}_t \to h(q)$ as $t \to \infty$ with $P_{\tilde{\pi}}(\cdot|X_0 = i, q)$-probability 1. Also, from (ii) of Lemma 5, $\tilde{q}_{ij}^t(a) \to q_{ij}(a)$ as $t \to \infty$ with $P_{\tilde{\pi}}(\cdot|X_0 = i, q)$-probability 1. So, observing Step 3 of Algorithm $(*)$, we have

$$\tilde{\pi}_t(A^*(j|q)|X_0, \Delta_0, \ldots, \Delta_{t-1}, X_t = j) = \tilde{\pi}(A^*(j|q)|j) \to 1 \quad \text{as } t \to \infty.$$

Thus, from Lemma 4, it follows that $\tilde{\pi}$ is q-optimal, which completes the proof. □

4 A numerical experiment

In this section, we consider the adaptive case of "Toymaker's problem" in Howard [7], for which a numerical experiment by TD-based learning algorithm (Algorithm $(*)$) given in Section 3 is put in practice. For the corresponding MDPs, there exists two states $S = \{1, 2\}$, where state 1(2) represents that he has a successful(unsuccessful) toy. In state 1(2), he has two actions, where action 1(1) is "no advertising"("no research") and action 2(2) is "advertising"("research"). The table of the unknown (true) stochastic matrix q and reward function r and the figure of transition diagrams are given Table 1 and Figure 1.

i	a	$q_{ij}(a)$		$r(i,a)$
		1	2	
1	1	0.5	0.5	6
	2	0.8	0.2	4
2	1	0.4	0.6	−3
	2	0.7	0.3	−5

Table 1. Toymaker's Problem

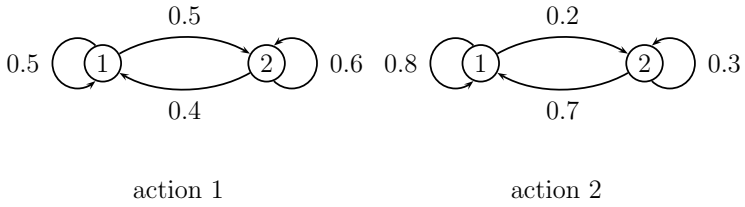

Fig. 1. Transition diagrams in Toymaker's problem.

By solving the optimality equation (3), we find that the optimal policy f^* is such that $f^*(1) = f^*(2) = 2$ and the optimal average reward $g^* = 2$ with $v(1) - v(2) = 10$.

values	δ \ t	10^3	5×10^3	10^4	5×10^4	10^5	10^6
$\tilde{g}_t$	0.1	1.436563	1.874825	1.884112	1.930601	1.971630	2.002304
	0.01	1.436563	1.872825	1.883312	1.930441	1.971550	2.002296
decision rule	δ \ t	10^3	5×10^3	10^4	5×10^4	10^5	10^6
$\tilde{\pi}_t(1\|1)$	0.1	0.998299	0.000275	0.000134	0.000026	0.000013	0.000001
$\tilde{\pi}_t(2\|1)$		0.001701	0.999725	0.999866	0.999974	0.999987	0.999999
$\tilde{\pi}_t(1\|2)$		0.002387	0.000729	0.000393	0.000086	0.000044	0.000005
$\tilde{\pi}_t(2\|2)$		0.997613	0.999271	0.999607	0.999914	0.999956	0.999995
$\tilde{\pi}_t(1\|1)$	0.01	0.998299	0.000275	0.000134	0.000026	0.000013	0.000001
$\tilde{\pi}_t(2\|1)$		0.001701	0.999725	0.999866	0.999974	0.999987	0.999999
$\tilde{\pi}_t(1\|2)$		0.002387	0.000730	0.000394	0.000086	0.000044	0.000005
$\tilde{\pi}_t(2\|2)$		0.997613	0.999270	0.999606	0.999914	0.999956	0.999995

Table 2. The simulation value of $\tilde{g}_t$ and $\tilde{\pi}_t$ for $\delta = 0.1, 0.01$

We denote by $\tilde{g}_t$ the average present value obtained from adaptive policy $\tilde{\pi} = (\tilde{\pi}_0, \tilde{\pi}_1, \ldots)$ constructed through Algorithm $(*)$, which is denoted by

$$\tilde{g}_t = \frac{1}{t}\sum_{l=0}^{t-1} r(X_l, \Delta_l) \quad (t \geqq 1). \tag{15}$$

We have simulated Algorithm $(*)$ to investigate the asymptotic behaviour of $\tilde{g}_t$ $(t \geqq 1)$ under the adaptive policy $\tilde{\pi}$. The data for simulation have been given as follows:

(i) The initial policy $\tilde{\pi}_0(\cdot|i) = \left(\frac{1}{2}, \frac{1}{2}\right), (i \in S)$ and $q^0 = \{(q^0_{ij}(a))|q^0_{ij}(a) = \frac{1}{2}$ for $i, j \in S, a \in A\}$,

(ii) $\varepsilon_t(i) = (N_t(i) + K(i) + 1)^{-1}, \gamma(t)(i) = (N_t(i)/1000 + 10)^{-1}$ $(t \geqq 1, i \in S)$, where $K(i)$ is given in (10) and $N_t(i)$ represents the number of visiting state i until the t-th time.

These data are shown to satisfy Condition $(*)$. The results of simulations with $\delta = 0.1, 0.01$ are given in Table 2 and Figure 2, in which we observe that

$$\tilde{\pi}_t(2|1) \to 1, \tilde{\pi}_t(2|2) \to 1 \ \text{ and } \ \tilde{g}_t \to g = 2 \ \text{ as } \ t \to \infty$$

These results surely show the efficiency of the adaptive policy $\tilde{\pi}$ constructed through TD-based learning algorithm(Algorithm $(*)$).

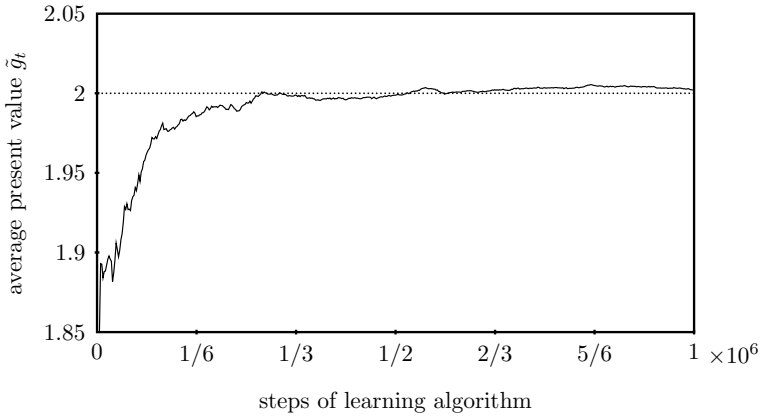

Fig. 2. The trajectories of $\tilde{g}_t(\delta = 1)$. The dotted line means the optimal value of average reward.

References

1. J. Abounadi, D. Bertsekas, and V. S. Borkar. Learning algorithms for Markov decision processes with average cost. *SIAM J. Control Optim.*, 40(3):681–698, 2001.
2. D. P. Bertsekas and J. H. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Massachusetts, Belmont, 1996.
3. P. Billingsley. *Statistical inference for Markov processes*. Statistical Research Monographs, Vol. II. The University of Chicago Press, Chicago, Ill., 1961.
4. V. S. Borkar and S. P. Meyn. The O.D.E. method for convergence of stochastic approximation and reinforcement learning. *SIAM J. Control Optim.*, 38(2):447–467, 2000.
5. A. Federgruen and P. J. Schweitzer. Nonstationary Markov decision problems with converging parameters. *J. Optim. Theory Appl.*, 34(2):207–241, 1981.
6. O. Hernández-Lerma. *Adaptive Markov control processes*, volume 79 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1989.
7. R. A. Howard. *Dynamic programming and Markov processes*. The Technology Press of M.I.T., Cambridge, Mass., 1960.
8. T. Iki, M. Horiguchi, and M. Kurano. A structured pattern matrix algorithm for multichain markov decision processes. *(To appear in MMOR)*, 2007.
9. T. Iki, M. Horiguchi, M. Yasuda, and M. Kurano. A learning algorithm for communicating markov decision processes with unknown transition matrices. *(submitted)*.
10. V. R. Konda and V. S. Borkar. Actor-critic-type learning algorithms for Markov decision processes. *SIAM J. Control Optim.*, 38(1):94–123 (electronic), 1999.
11. M. Kurano. Adaptive policies in Markov decision processes with uncertain transition matrices. *J. Inform. Optim. Sci.*, 4(1):21–40, 1983.
12. M. Kurano. Learning algorithms for Markov decision processes. *J. Appl. Probab.*, 24(1):270–276, 1987.
13. A. Leizarowitz. An algorithm to identify and compute average optimal policies in multichain Markov decision processes. *Math. Oper. Res.*, 28(3):553–586, 2003.

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. E. Nummelin. *General irreducible Markov chains and nonnegative operators*, volume 83 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1984.
17. M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons Inc., New York, 1994. A Wiley-Interscience Publication.
18. R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 1998.