Adaptive Markov decision processes based on temporal difference method

伊喜哲一郎　（宮崎大学教育文化学部）　堀口正之　（弓削商船高専総合教育科）
安田正實　（千葉大学理学部）　蔵野正美　（千葉大学教育学部）

We consider an adaptive model for uncertain Markov decision processes(MDPs) with the average reward criterion. In our previous work [2], the adaptive policies are constructed by applying the methods of value iteration, cooperated with the policy improvement(cf. [4]), in which the corresponding value function is approximated through the expectation with respect to the estimated transition matrices at each learning step. In this paper, under the minorization condition, we construct an adaptive policy based on temporal difference method in neuro-dynamic programming. By the stochastic iteration algorithm, the estimated value function is updated using temporal difference and adaptive policy are constructed as an ε-forced modification of the greedy policy for the estimates of value function and transition probability matrices.

Key words: Adaptive Markov decision processes, neuro-dynamic programming, temporal difference, average case.

1. Introduction and notation

Consider a controlled dynamic system with finite state and action spaces, S and A, which consist of finite N and K elements respectively. For any $\delta > 0$, let $\mathbb{Q}_\delta$ denote the parameter space of K unknown stochastic matrices, defined by

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

A transition matrix of this class $\mathbb{Q}_\delta$ is the minorization condition (named after [3]) and is assumed throughout the paper. 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 randomized stationary policy ξ and a stationary policy f are defined by a usual way (cf. [2, 4]).

We will construct a probability space as follows: for any initial state $X_0 = i, \pi \in \Pi$ and a transition law $q = (q_{ij}(a))$, 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:

$$\psi(i, q|\pi) = \liminf_{T\to\infty} \tfrac{1}{T+1} E_\pi\big(\textstyle\sum_{t=0}^{T} r(X_t, \Delta_t) \;\big|\; X_0 = i, q\big) \tag{2}$$

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.,

$$\psi(i, q) = \sup_{\pi\in\Pi} \psi(i, q|\pi), \tag{3}$$

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

2. Preliminary lemmas

Let $B(S)$ be the set of all functions on S. For any $q \in \mathbb{Q}_\delta$, we define the map $U\{q\} : B(S) \to B(S)$ by

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

Let $h(q) \in B(S)$ be unique fixed point of $U(q)$.

Lemma 1. *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 $\psi^*(q) = \delta \sum_{j\in S} h(q)(j)$ and $A^*(i|q)$ is the set of optimal actions at state i and $A^*(i|q) = \arg\max_{a\in A}\big\{r(i, a) - \psi^*(q) + \sum_{j\in S} q_{ij}(a)h(q)(j)\big\}$.*

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

$$\tilde{v}_0(i) \equiv 0, \tilde{v}_{t+1}(i) = \big(1 - \tilde{\gamma}_t(i)\big)\tilde{v}_t(i) + \tilde{\gamma}_t(i)\Big(H\tilde{v}_t(i) + W_t(i) + u_t(i)\Big), \quad t \geqq 0 \tag{5}$$

where $\tilde{\gamma}_t(i)$ is 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 $i \in S$.

Lemma 2 (cf. Proposition 4.5 in [1]). *Suppose that the following condition (i) – (v) are assumed to hold. (i) $E[W_t(i)|\mathcal{F}_t] = 0$ for $i \in S$ (ii) There exist $A, B > 0$ such that $E\big[W_t(i)^2 \mid \mathcal{F}_t\big] \leqq A + B\|\tilde{v}_t\|^2$for $t \geqq 0$ and $i \in S$. (iii) H is a contraction operator 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 $\{\theta_t\}$ such that $|u_t(i)| \leqq \theta_t(||\tilde{v}_t|| + 1)$ for $i \in S$ and $t \geqq 0$ and $\{\theta_t\}$ converges to zero with probability 1. Then, $\tilde{v}_t$ in (5) 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. Temporal difference-based adaptive policies

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. For any given $q^0 = (q_{ij}^0(a)) \in \mathbb{Q}_\delta$, we define $\tilde{q}^n = (\tilde{q}_{ij}^n(a)) \in \mathbb{Q}_\delta$ by $\tilde{q}_{ij}^n(a) = q_{ij}^n(a)$ if $N_n(i|a) > 0$, $= q_{ij}^0(a)$ otherwise. 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 $(*)$:

step 1. Set $t = 0$ and $\tilde{v}_0 \equiv 0$ and let $\tilde{\pi}_0 \in P(A|S)$ with $\tilde{\pi}_0(a|i) > 0$ for all $a \in A$ and $i \in S$.

step 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 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 $i \in S$,

$$\tilde{v}_{t+1}(i) = \big(1 - \tilde{\gamma}_t(i)\big)\tilde{v}_t(i) + \tilde{\gamma}_t(i)\big(r(i, \Delta_t) + \tilde{v}_t(X_{t+1}) - \delta\textstyle\sum_{\ell \in S}\tilde{v}_t(\ell)\big) \tag{6}$$

where the step size $\tilde{\gamma}_t$ is defined as in (5) for a sequence $\{\gamma_t(i)\}$.

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

$$\tilde{\pi}_{t+1}(a|i) = \varepsilon_t(i)/(K(i) - 1) \text{ if } a \neq \tilde{a}_{t+1}(i), = 1 - \varepsilon_t(i) \text{ if } a = \tilde{a}_{t+1}(i), \tag{7}$$

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

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

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 $i \in S$.

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.*

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

4. A numerical experiment

References

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

[2] 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).*

[3] Esa Nummelin. *General irreducible Markov chains and nonnegative operators*, volume 83 of *Cambridge Tracts in Mathematics.* Cambridge University Press, Cambridge, 1984.

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

References

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

[2] 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).*

[3] Esa Nummelin. *General irreducible Markov chains and nonnegative operators*, volume 83 of *Cambridge Tracts in Mathematics.* Cambridge University Press, Cambridge, 1984.

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