Adaptive algorithm for Markov decision processes using pattern-matrix learning method

伊喜哲一郎 （宮崎大学教育文化学部） 堀口正之 （神奈川大学工学部）
安田正實 （千葉大学理学部） 蔵野正美

This note is concerned with adaptive algorithm for uncertain Markov decision processes(MDPs) with the average reward criterion. As a sequel to [2], we consider the learning algorithm for investigating the structure of unknown transition matrices and getting adaptive policy in regularly communicating model where the state space is decomposed into a single communicating class and a transient class. Using pattern matrix learning method, we have an asymptotic sequence of adaptive policies with nearly average optimal properties.

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

1. Notation

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.,

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

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 ($t \geqq 0$). 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. [3]).

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. $H_{n-1} = (X_0, \Delta_0, \ldots, X_{n-1})$ denotes a history until the $(n-1)$-th step. Let $n(D)$ denotes the number of elements in a set D.

For a given reward function r on $S \times A$, we shall consider the long-run expected average reward:

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

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

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

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

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 $q_{i_1 i_2}(a_1) q_{i_2 i_3}(a_2) \cdots q_{i_{l-1} i_l}(a_{l-1}) > 0$ 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., $P_\pi(X_t \in \bar{E} \text{ for some } t \geqq 1 | X_0 \in T) = 1$ 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 $i_0 \in \bar{E}(q)$.

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 [2].

2. Preliminary lemmas

Lemma 1. *Let $q \in \mathbb{Q}^*(i_0)$ with $i_0 \in S$. Let a policy $\tilde{\pi} = (\tilde{\pi}_0, \tilde{\pi}_1, \ldots)$ and a decreasing sequence of positive numbers $\{\varepsilon_t\}_{t=0}^{\infty}$ satisfy that for each $t \geqq 0$ $\tilde{\pi}_t(a|h_t) \geqq \varepsilon_t$ 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 $E \subsetneqq \bar{E}(q)$,*

$$P_{\tilde{\pi}}(X_{t+l} \in \bar{E}(q) - E \text{ for some } l(1 \leqq l \leqq N)|X_t \in E) \geqq (\delta\varepsilon_{t+N})^N. \tag{4}$$

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:

$$\sigma_n := \min\{t|X_t \in T_{\sigma_{n-1}}, t > \sigma_{n-1}\}, E_{\sigma_n} = E_{\sigma_{n-1}} \cup \{X_{\sigma_n}\}, T_{\sigma_n} = \bar{E}(q) - E_{\sigma_n},\ (n \geqq 1) \tag{5}$$

where $\min \emptyset = \infty, \sigma_0 = 0, E_0 := \{i_0\}$ and $T_0 := \bar{E}(q) - E_0$.

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

3. Pattern-matrix learning algorithms

For any sequence $\{b_n\}_{n=0}^{\infty}$ 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 $\mathbf{PMLA}(\boldsymbol{i_0}, \{\boldsymbol{b_n}\}, \boldsymbol{\tau})$(cf. [2]).

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

1. Set $E_0 = \{i_0\}, T_0 = S - E_0, \tilde{v}_0(i_0) = 0$, $X_0 = i_0$ and $\tilde{\pi}_0^{\tau}(a|X_0) = n(A(i_0))^{-1}$ for $a \in A(i_0)$.
2. Suppose that $E_n \subset S, T_n = S - E_n$, $\{\tilde{v}_n(i)\}$, $\tilde{\pi}_n^{\tau}(a|i) = Prob.(\Delta_n = a|H_{n-1}, \Delta_{n-1}, X_n = i)$ $(i \in E_n, a \in A(i))$ are given.
3. Choose $\Delta_{n+1} \in A(X_n)$ from $\tilde{\pi}_n(\cdot|H_n)$. 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
$$q_{ij}^{n+1}(a) = \frac{N_{n+1}(i,j|a)}{N_{n+1}(i,a)} \text{ if } N_{n+1}(i|a) > 0,\ q_j^0 \text{ otherwise },\ (i,j \in E_{n+1}, a \in A(i)) \tag{6}$$
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
$$\tilde{a}_{n+1}(i) \in \arg\max_{a \in A(i)} \{r(i,a) + (1-\tau) \textstyle\sum_{j \in E_{n+1}} q_{ij}^{n+1}(a) \tilde{v}_n(j)\}$$
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:
$$\tilde{\pi}_{n+1}^{\tau}(a_i|i) = 1 - \textstyle\sum_{a \neq a_i} \phi(\tilde{\pi}_n^{\tau}(a|i)) \text{ if } a_i = \tilde{a}_{n+1}(i)),\ = \phi(\tilde{\pi}_n^{\tau}(a_i|i)) \text{ if } a_i \neq \tilde{a}_{n+1}(i)\ . \tag{7}$$
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 $(*)$ $b_n \to 0$ as $n \to \infty$ and $\sum_{n=0}^{\infty} b_n^N = \infty$.

Combining the vanishing discount approach(cf. [2, 3]) as letting $\tau \to 0$ with $\mathbf{PMLA}(\boldsymbol{i_0}, \{\boldsymbol{b_n}\}, \boldsymbol{\tau})$, we have the policy $\tilde{\pi}^{\tau} = (\tilde{\pi}_0^{\tau}, \ldots)$ which has nearly average-optimal properties for $\mathbb{Q}^*(i_0)$ as follows:

Theorem 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 $\mathbb{Q}^*(i_0)$.*

4. A numerical experiment

References

[1] T. Iki, M. Horiguchi, M. Yasuda, and M. Kurano. Adaptive Markov decision processes based on temporal difference method. 日本数学会 *2007* 年度秋季 統計数学分科会講演アブストラクト, 125–126, 2007.

[2] T. Iki, M. Horiguchi, M. Yasuda, and M. Kurano. A learning algorithm for communicating Markov decision processes with unknown transition matrices. *Bulletin of Information and Cybernetics*, 39:11–24, 2007.

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