マルコフ決定過程における適応型アルゴリズム

堀口正之*

Adaptive Algorithms for Markov Decision Processes

Masayuki HORIGUCHI*

1. はじめに

マルコフ決定過程は, 意思決定と確率的な状態推移を表現する多段決定過程の一つであり, その端緒を開いたのはBellman[4]である. "マルコフ連鎖"や"動的計画法"のキーワードは今日では馴染み深いものである. Bellman[3, 4]やHoward[9]による著書は多段決定過程としてモデル化できる問題を扱う幅広い分野の研究発展に影響を与えた. マルコフ決定過程はその表現される動的システムの構造から統計科学や確率的な最適制御, 適応制御などの問題として定式化され, コンピュータの進歩も伴い動的計画法のいわゆる「次元の呪い(Curse of Dimensionality)」の克服に道が開け人工知能の研究分野でも「強化学習(Reinforcement Learning)」と呼ばれるニューロ・ダイナミックプログラミングによる学習アルゴリズムの研究が活発になされている(cf. [1, 5, 6, 13, 24]).

本稿では, 不確実性の下でのマルコフ決定過程の問題において推移確率法則が未知であるモデルの適応型学習理論について述べる. 適応型マルコフ決定過程の先行研究としては[8, 14, 17, 18]などが上げられる. ここで扱うモデルは部分観測可能なマルコフ決定過程(e.g. [20])の問題とは異なり, 各期の状態は観測できるが状態推移確率が未知であるため政策反復法や価値反復法による最適政策の解析はできない. 適切な条件下において, 1. 状態集合がただ一つの互いに到達可能な状態から成る場合, 2. 状態集合が一つの互いに到達可能な状態類と一つの絶対消散状態類から成る場合, 3. 状態集合がマイノリゼーション条件を満たすときのニューロ・ダイナミックプログラミングによる時間差分法の適用, について逐次, 状態を観測し推移法則に関する情報を収集しながら適応的に最適政策を構築する学習アルゴリズムをそれぞれ示す.

2. 準備

マルコフ決定過程(Markov Decision Processes, MDPs)は, 次の4つの構成要素によって定義される:

$$\{S, A, (q_{ij}(a)), r(\cdot,\cdot)\}.$$

S, A はそれぞれ状態空間(state space)と決定空間(action space)を表し本報告では S, A の要素の数は有限

$$|S| = N < \infty, |A| = K < \infty$$

であることを仮定する. $q_{ij}(a)$ は意思決定者(decision maker)が状態 i で決定 a と取ったとき次の期に状態 j へ推移する確率を表す. また, $r(i,a)$ $(i \in S, a \in A)$ は状態 i で決定 a を取ったときの利得関数を表す.

マルコフ決定過程における基本的な最適化問題においては, 完全情報とくに推移確率が既知であることを前提にしているが, ここでは推移確率が未知の問題(Uncertain MDPs)について考察する. K 個の未知の推移確率行列によるパラメータ空間を

$$\mathbb{Q} = \Big\{q \mid q = (q_{ij}(a) : i, j \in S, a \in A), \quad q_{ij}(a) \geqq 0, \sum_{j \in S} q_{ij}(a) = 1 \ (i, j \in S, a \in A)\Big\}$$

と定義する.

記号 $\pi = (\pi_0, \pi_1, \ldots)$ は政策(policy)を表す. 政策の全体集合を Π と表す. 例えば, 第 n 期に状態 i_n で決定 a_n をとる確率が p_n であればその時のマルコフ政策は

$$\pi_n(a_n|i_n) = p_n$$

と表される. 特に, それぞれの状態 i で特定の決定 a_i を確定的に取る政策を定常政策と呼び, $f : S \to A$ で $f(i) = a_i$ である関数 f として表現する.

$$\pi_n(f(i)|i) = 1 \ (i \in S, n \geqq 0)$$

であり $f \in \Pi$ と表される. 一般に第 n 期の政策は

$$\pi_n \in P(A|(S \times A)^n \times S) \ (n \geqq 0)$$

と条件付き確率で表される. また, 第 n 期までの履歴(history)を表す確率変数を

$$H_0 := X_0, H_n := (X_0, \Delta_0, X_1, \ldots, X_n) \ (n \geqq 1)$$

と表す. ただし X_n, Δ_n はそれぞれ第 n 期の状態と決定を表す確率変数である. マルコフ決定過程に関する詳しい用語の定義や取り扱いについては例えば[22, 28]などを参照されたい.

期待平均利得(long-run expected average reward)関

*准教授　数学教室
Associate Professor, Institute of Mathematics

数を

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

とする．ただし，$E_\pi(\cdot|X_0 = i, q)$ は初期状態が i で $q \in \mathbb{Q}$ のときの政策 π による確率測度 $P_\pi(\cdot|X_0 = i, q)$ に関する期待値を表す．$\mathcal{D}$ を $\mathbb{Q}$ の部分集合とする．すべての政策 $\pi \in \Pi$ と任意の $i \in S, q \in \mathcal{D}$ について式(1)を最大化する最適化問題を考える，すなわち，

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

を価値関数(value function)とし，すべての状態 $i \in S$ について $\psi(i,q|\pi^*) = \psi(i,q)$ が成り立つ政策 $\pi^* \in \Pi$ を q-最適(q-optimal)政策と呼び，すべての $q \in \mathcal{D}$ に対して π^* が q-最適であるとき，その π^* を $\mathcal{D}$ に関して適応型最適政策(adaptively optimal)であると呼ぶ．

各期の政策列 $\{\pi_n\}_{n=0}^{\infty}$ について

$$\lim_{n\to\infty} \psi(i,q|\pi_n) = \psi(i,q) \quad \text{for all } q \in \mathcal{D}$$

であるとき，この政策列をほとんど最適な適応型政策列(asymptotic sequence of adaptive policies with nearly optimal properties)と呼ぶ．

以後，推移確率行列 $\mathbb{Q}$ を3つの場合に分けてそれぞれの適応型政策を構成するためのアルゴリズムについて考察する．必要な用語の定義といくつかの命題をこの節でまとめておく．

• 到達可能行列: 推移確率行列 $q = (q_{ij}(a)) \in \mathbb{Q}$ について，任意の状態 $i, j \in S$ が互いに到達可能(communicating)であるとき，すなわち，ある $\{i_1 = i, i_2, \ldots, i_l = j\} \subset S, \{a_1, a_2, \ldots, a_{l-1}\} \subset A, 2 \leqq l \leqq N$ について

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

が成り立つとき，i から j へ到達可能であると言い $i \to j$ と表し q を到達可能行列と呼ぶ．到達可能行列の全体を $\mathbb{Q}^*$ で表す．

• 到達可能類: 状態部分集合 $E \subset S$ について，$q \in \mathbb{Q}$ が次の条件を満たすとき，E を q の到達可能類(communicating class)と呼ぶ．

(i) 任意の状態 $i, j \in E$ について $i \to j$，

(ii) E は閉じている，すなわち，

$$\sum_{j\in E} q_{ij}(a) = 1 \ (i \in E, a \in A(i)).$$

ただし，$A(i)$ は状態 i での選択可能な決定(available actions)の集合を表す．

• 正規到達可能行列: 推移確率行列 $q \in \mathbb{Q}$ について，ある状態部分集合 $\bar{E} \subsetneqq S$ が存在して，

(i) $\bar{E}$ は q の到達可能類である，

(ii) $T = S - \bar{E}$ は一つの絶対消散類(absolutely transient class)である，すなわち，すべての $\pi \in \Pi$ に対して

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

が成り立つとき，q を正規到達可能(regularly communicating)行列と呼ぶ．正規到達可能行列 q に応じて決まる上記(i)の到達可能類 $\bar{E}$ を $\bar{E}(q)$ で表す．状態 $i_0 \in S$ が $i_0 \in \bar{E}(q)$ となるような正規到達可能行列の集合を $q \in \mathbb{Q}^*(i_0)$ と表す．

• マイノリゼーション条件: 任意の $\delta > 0$ を一つ選び固定したとき

$$\mathbb{Q}_\delta = \Big\{ q = (q_{ij}(a)) \,\Big|\, q_{ij}(a) \geqq \delta, \sum_{j\in S} q_{ij}(a) = 1 \quad \text{for } i, j \in S, a \in A \Big\} \quad (2)$$

をマイノリゼーション条件を満たす行列集合と呼ぶ．

• 割引消滅法: 割引率 $(1-\tau)$ の $i \in S, q \in \mathbb{Q}, \pi \in \Pi$ についての期待総利得(expected total $(1-\tau)$-discount reward)を

$$v_\tau(i,q|\pi) = E_\pi \left(\sum_{t=0}^{\infty} (1-\tau)^t r(X_t, \Delta_t) \,\middle|\, X_0 = i, q \right) \quad (3)$$

と定義する．また，

$$v_\tau(i,q) = \sup_{\pi\in\Pi} v_\tau(i,q|\pi)$$

を割引率 $(1-\tau)$ の価値関数と呼ぶことにする．S 上のすべての関数の全体集合を $B(S)$ と表す，すなわち，

$$B(S) = \{f | f : S \to R\}$$

とおく．$q = (q_{ij}(a)) \in \mathbb{Q}$ と $\tau \in (0,1)$ に対して作用素 $U_\tau\{q\} : B(S) \to B(S)$ を各 $i \in S, u \in B(S)$に関して次のように定義する:

$$U_\tau\{q\}u(i) = \max_{a\in A} \Big\{ r(i,a) + (1-\tau) \sum_{j\in S} q_{ij}(a) u(j) \Big\}. \quad (4)$$

次の補題は割引消滅法(vanishing discount approach)として知られている．(cf. [22, 23]).

補題 1.

(i) 作用素 $U_\tau\{q\}$ は係数 $(1-\tau)$ を持つ縮小写像である．

(ii) 割引率 $(1-\tau)$ の目的関数 $v_\tau(i,q)$ は作用素 $U_\tau\{q\}$ の一意な不動点である．すなわち，

$$v_\tau = U_\tau\{q\} v_\tau$$

を満たす．

(iii) f_τ を式(4)の右辺の最大化(maximizer)関数とするとき,

$$v_\tau(i,q) = v_\tau(i,q|f_\tau),\ \lim_{\tau\to 0}\tau v_\tau(i,q) = \psi(i,q)$$

が成り立つ.

はじめに, 到達可能行列 $q \in \mathbb{Q}^*$ に関する最適方程式を導く. $\mu = (\mu_1, \mu_2, \ldots, \mu_N) \in P(S)$ を用いて, q に関する摂動推移確率行列 $q^{\tau,\mu} = (q_{ij}^{\tau,\mu}(a))$ を次のように定義する: $i,j \in S, a \in A$ のそれぞれについて

$$q_{ij}^{\tau,\mu}(a) = \tau\mu_j + (1-\tau)q_{ij}(a). \tag{5}$$

式(5)は N-次元列ベクトル $(1,1,\ldots,1)$ の転置ベクトル $e = (1,1,\ldots,1)^t$ を用いて

$$q^{\tau,\mu} = \tau e\mu + (1-\tau)q$$

と行列で表せる. $P(S)$ を S 上の確率分布とする.

定理 1. ($\mathbb{Q}^*$ での最適方程式) 到達可能行列 $q \in \mathbb{Q}^*$ を任意に一つ選び固定する. このとき

(i) $\psi(i,q)(:=\psi(q))$ は状態 $i \in S$ に依存しない.さらにある関数 $u \in B(S)$ は次の最適方程式を満たす: $i \in S$ について

$$u(i) = \max_{a\in A}\Big\{r(i,a) + \sum_{j\in S} q_{ij}(a)u(j)\Big\} - \psi(q) \tag{6}$$

(ii) 任意の $\mu \in P(S)$ に対して $\tau \to 0$ とするとき $\psi(q^{\tau,\mu}) \to \psi(q)$.

次に $\mathbb{Q}_\delta$ に関する最適方程式を示そう. マイノリゼーション条件を満たす推移確率行列 $q \in \mathbb{Q}_\delta$ に対して作用素 $U\{q\} : B(S) \to B(S)$ を次のように定義する:

$$U\{q\}u(i) = \max_{a\in A}\Big\{r(i,a) + \sum_{j\in S}\big(q_{ij}(a) - \delta\big)u(j)\Big\} \tag{7}$$

このとき, $U(q)$ が縮小写像であることは容易に示される. $h(q) \in B(S)$ を $U(q)$ の一意の不動点とする, すなわち,

$$h(q) = U\{q\}h(q)\ (q \in \mathbb{Q}_\delta) \tag{8}$$

であるとするとき, 式(8)において

$$\psi^*(q) = \delta\sum_{j\in S} h(q)(j)$$

とすれば

$$h(q)(i) = \max_{a\in A}\Big\{r(i,a) - \psi^*(q) + \sum_{j\in S} q_{ij}(a)h(q)(j)\Big\} \tag{9}$$

と平均期待利得に関する最適方程式として表されることから次の定理を得る:

定理 2. ($q \in \mathbb{Q}_\delta$ に関する最適政策)

$$\psi^*(q) = \psi(i,q)\ (i \in S)$$

である. すなわち, $\psi(i,q)$ の値は初期状態に依存しない. さらに, もし, すべての $i \in S$ に対して $f(i) \in A^*(i|q)$ とすれば f は q 最適定常政策である. ただし $A^*(i|q)$ は状態 i での最適決定(optimal action)の集合で式(9)の右辺のmaximizerの決定集合である.

3. 適応型政策の構成

未知の推移確率行列を持つマルコフ決定過程での適応型政策の構成方法について説明しよう. 第 n 期に状態 $i \in S$ で決定 $a \in A(i)$ を取ったときの次の期の状態が $j \in S$ である頻度を次のように定義する:

$$N_n(i,j|a) = \sum_{t=0}^{n} I_{\{X_t=i,\Delta_t=a,X_{t+1}=j\}}, \tag{10}$$

$$N_n(i|a) = \sum_{t=0}^{n} I_{\{X_t=i,\Delta_t=a\}}. \tag{11}$$

ただし I_D は集合 D の指示関数である. $q_{ij}^n(a)$ を次のように定義する.

$$q_{ij}^n(a) := \begin{cases} \dfrac{N_n(i,j|a)}{N_n(i|a)}, & N_n(i|a) > 0 \text{ のとき,} \\ 0, & \text{その他.} \end{cases}$$

このとき $q_{ij}^n = (q_{ij}^n(a))$ は未知の推移確率行列に対する最尤推定量を表している. $q^0 = (q_{ij}^0(a)) \in \mathbb{Q}$ を任意に選び $\tilde{q}^n = (\tilde{q}_{ij}^n(a)) \in \mathbb{Q}$ を

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

によって定義する. 任意の $\tau \in (0,1)$ に対して更新関数(update function) $\{\tilde{v}_n\}_{n=0}^\infty$ に関する次のような反復スキーム(iterative scheme)を考える.

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

式(12)の第2の式における右辺の最大化関数のひとつを $\tilde{a}_{n+1}$ と表し, 各状態 i における最大化関数で定義される決定政策は $\tilde{a}_{n+1}(i)$ と表される. すなわち,

$$\tilde{a}_{n+1}(i) \in \arg\max_{a\in A(i)}\{r(i,a) + (1-\tau)\sum_{j\in E_{n+1}} \tilde{q}_{ij}^{n+1}(a)\tilde{v}_n(j)\}. \tag{13}$$

$\{b_n\}_{n=0}^\infty$ は $b_0 = 1$ である正の狭義単調減少数列であるとし $\phi : [0,1] \to [0,1]$ は

$$\phi(b_n) = b_{n+1}\ (n \geqq 0) \tag{14}$$

を満たす狭義単調増加関数とする. 適応型政策を構成するための学習アルゴリズム(learning algorithm)は式(12)の最大化関数 a_{n+1} と関数 ϕ から成る.

$$\tilde{\pi}_n^\tau(a|i) = P(\Delta_n = a|X_0, \Delta_0, \ldots, X_n = i)$$

とおくとき, 各反復での適応型政策 $\tilde{\pi}_n^\tau$ は次のように改定される: $a_i = \tilde{a}_{n+1}(i)$ $(i \in S)$ とするとき,

$$\begin{aligned} \tilde{\pi}_{n+1}^\tau(a_i|i) &= 1 - \sum_{a \neq a_i} \phi(\tilde{\pi}_n^\tau(a|i)), \\ \tilde{\pi}_{n+1}^\tau(a|i) &= \phi(\tilde{\pi}_n^\tau(a|i)) \ (a \neq a_i). \end{aligned} \tag{15}$$

政策列 $\tilde{\pi}^\tau = (\tilde{\pi}_0^\tau, \tilde{\pi}_1^\tau, \ldots)$ は初期政策 $\tilde{\pi}_0$ を与え, $\tilde{\pi}_n^\tau$ $(n \geqq 1)$ を式(12)と(15)によって逐次決める.

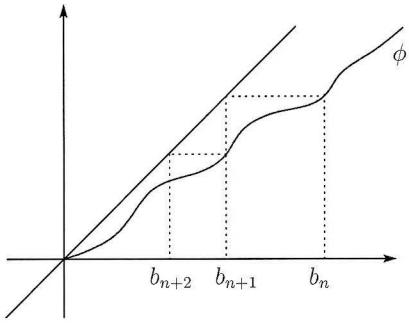

図1 ϕ と $\{b_n\}_{n=0}^\infty$ の関係

ここでいくつかの点について言及しておく. 式(12)は Federgruen and Schweitzer[7]によって提案された非定常価値反復スキーム(non-stationary value iteration scheme)の一つの変形ととらえることができる. さらに, 式(12)での更新価値関数 $\tilde{v}_n$ に関して最大化関数 $\tilde{a}_{n+1}$ を決めることは, その時点で得られている推移と更新関数の値の情報から一期間先に推移した時の価値を最大にするような貪欲政策(greedy action)を選択することであって, 式(15)は関数 ϕ によって次の期に取るべき政策は $\tilde{a}_{n+1}(i)$ の選択される確率が増加され他の決定の選択される確率は減少されるように更新される. また, ϕ の性質について詳しく説明すると, 正の狭義単調減少数列 $\{b_n\}$ で式(14)の関係を満たすような関数は, 直交座標平面での直線 $y = x$ $(x \geqq 0)$ のグラフの下側に関数 ϕ があって図1のように直線 $y = x$ 上に点 $(b_{n+1}, \phi(b_n))$ が取れることと視覚的に確認できる. 式(15)の学習アルゴリズムは利得罰金(reward-penarty)タイプ[15]と呼ばれる(cf. [16, 19, 24]). また, 到達可能行列と正規到達可能行列における以下の結果はKurano[15]における推移確率行列が

$$\begin{aligned} \mathbb{Q}^+ := \{q = (q_{ij}(a)) \in \mathbb{Q} | \\ q_{ij}(a) > 0 \text{ for all } i, j \in S \text{ and } a \in A\} \end{aligned} \tag{16}$$

で与えられる適応型最適政策の結果の拡張であり本報告で述べられる定理の証明方法は一部その結果の適用によっている.

3.1 到達可能行列

反復スキームにおける $\tilde{q}^n, \tilde{v}_n, \tilde{\pi}_n^\tau$ に関するそれぞれの収束性を得るために次の仮定をする:

仮定 1.

(i) $b_n \to 0$ $(n \to \infty)$, $\sum_{n=0}^{\infty} b_n^N = \infty$.

(ii) すべての $i \in S, a \in A$ に対して $\tilde{\pi}_0^\tau(a|i) > 0$.

このとき次のような補題が得られる.

補題 2. $q \in \mathbb{Q}^*$ とする. 仮定 1のもとで以下の (i)–(iii) が $P_{\tilde{\pi}^\tau}(\cdot|X_0 = i, q)$-a.s. で成り立つ:

(i) $\tilde{q}^n \to q$ $(n \to \infty)$,

(ii) $\tilde{v}_n(i) \to v_\tau(i, q)$ $(n \to \infty)$,

(iii) $\tilde{\pi}_n^\tau(A_\tau^*(i|q)|H_n, X_n = i) \to 1$ $(n \to \infty)$,

ただし $A_\tau^*(i|q)$ は式(6)の右辺のmaximizerの集合である.

${}^\tau\mathbb{Q}^* := \{q^{\tau,\mu} | \mu \in P(S), q \in \mathbb{Q}^*\}$とおく. 摂動推移確率行列の族 ${}^\tau\mathbb{Q}^*$ について次の定理が得られる:

定理 3. 仮定 1のもとで $\tilde{\pi}^\tau$ は ${}^\tau\mathbb{Q}^*$ 関して適応型最適政策である.

ここでは詳しく取り上げないが, $\tau_n \to 0$ $(n \to \infty)$ とすることによる価値関数 v_{τ_n} の連続性と摂動行列 $q^{\tau_n,\mu}$ の連続性などにより, 反復スキームを進めていくことで得られる政策列が漸近的に最適政策となるほとんど最適な適応型政策を得ることが示される[11]:

定理 4. 仮定 1 のもとで, 政策列 $\{\tilde{\pi}^{\tau_n}\}_{n=1}^\infty$ $(\tau_n \to 0 \text{ as } n \to \infty)$ は $\mathbb{Q}^*$ に関するほとんど最適な適応型政策列である.

3.2 正規到達可能行列

次に正規到達可能行列 $\mathbb{Q}^*(i_0)$ での最適政策の存在に関して説明する. $E \subsetneqq \bar{E}(q)$ を取り以下のような状態部分集合列 $J_k(E)$ $(k = 1, 2, \ldots)$ を再帰的に定義する:

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

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

また,

$$\begin{aligned} K(\bar{E}(q)) = \{(i, a, j) | p_{ij}(a) > 0, \\ i, j \in \bar{E}(q), a \in A(i)\} \end{aligned}$$

と定義し, この $K(\bar{E}(q))$ に関する推移確率行列 q の成分の最小値を δ とおく, すなわち,

$$\delta = \min_{(i,a,j) \in K(\bar{E}(q))} p_{ij}(a)$$

とおく．このとき，以下の補題が示される．

補題 3. 任意の $q \in \mathbb{Q}^*(i_0)$ $(i_0 \in S)$ と $E \subsetneqq \bar{E}(q)$ を選ぶとき，ある自然数 $l(E)$ $(1 \leqq l(E) \leqq N)$ で

$$J_k(E) \neq \emptyset \ (k = 1, 2, \ldots, l(E)), J_{l(E)+1}(E) = \emptyset$$

を満たすものが存在する．

補題 4. $q \in \mathbb{Q}^*(i_0)$ $(i_0 \in S)$ とする．政策 $\tilde{\pi} = (\tilde{\pi}_0, \tilde{\pi}_1, \ldots)$ と正の実数の単調減少列 $\{\varepsilon_n\}_{n=0}^{\infty}$ は，各 $n \geqq 0$ について $\tilde{\pi}_n(a|h_n) \geqq \varepsilon_n$ $(a \in A(x_n), h_n = (x_0, a_0, x_1, \ldots, x_n) \in H_n)$ とする．このとき任意の状態部分集合 $E \subsetneqq \bar{E}(q)$ に対して

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

が成り立つ．

状態 $i_0 \in S$ について $q \in \mathbb{Q}^*(i_0)$ を一つ選んだとき，停止時刻列 $\{\sigma_n\}$ とそれによって定められる部分集合列 $\{E_{\sigma_n}\} \subset \bar{E}(q)$ を次のように定義する:

$$\begin{aligned} &E_0 := \{i_0\}, T_0 := \bar{E}(q) - E_0, \\ &\sigma_1 := \min\{t|X_t \in T_0, t > 0\}, \\ &E_{\sigma_1} = E_0 \cup \{X_{\sigma_1}\}, T_{\sigma_1} := \bar{E}(q) - E_{\sigma_1}, \\ &\text{とおき以下再帰的に } n = 2, 3, \ldots, \text{ について} \\ &\sigma_n := \min\{t|X_t \in T_{\sigma_{n-1}}, t > \sigma_{n-1}\}, \\ &E_{\sigma_n} = E_{\sigma_{n-1}} \cup \{X_{\sigma_n}\}, T_{\sigma_n} = \bar{E}(q) - E_{\sigma_n} \end{aligned} \tag{17}$$

と定義する．ただし $\min\emptyset = \infty$ である．

$E \subset \bar{E}(q)$ に対して

$$\bar{n}(E) = \min\{n \geqq 1|E_{\sigma_n} = \bar{E}(q)\}$$

とおく．このとき，もし $\bar{n}(E) < \infty$ ならば推移確率行列 q に関するパターン行列(pattern-matrix) $M(q)$ を構成することができる(cf. [10])．パターン行列は多重マルコフ連鎖(multi-cahin)における到達可能部分集合類を探索する際に用いられる行列表現であり，正規到達可能行列 $q \in \mathbb{Q}^*(i_0)$ に対しては

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

と表される．ここで $\bar{E}(q)$ の要素数を $n(\bar{E}(q))$ と表したとき E は $n(\bar{E}(q))$次正方小行列で，T は $n(S - \bar{E}(q)) \times n(\bar{E}(q))$ 小行列，E と R 成分はすべて1であって $M(q)$ において $i \to j$ は (i, j) 成分が1であることに相当する．正規到達可能行列の状態集合は小行列 E に対応する1つの到達可能類と T の対応する1つの絶対消散類に分類される．したがって，式(17)によって $\bar{n}(E) < \infty$ となることがわかればパターン行列 $M(q)$ を構成することができて正規到達可能行列に関する適応型政策の学習アルゴリズムは到達可能行列に関する問題に帰着できる．

補題 5. 推移確率行列 q を $q \in \mathbb{Q}^*(i_0)$ $(i_0 \in S)$ とする．$\tilde{\pi}$ は補題 4の仮定を $\sum_{t=0}^{\infty} \varepsilon_t^N = \infty$ となる $\{\varepsilon_t\}_{t=0}^{\infty}$ について満たしているとする．このとき任意の $E \subsetneqq \bar{E}(q)$ に対して次が成り立つ:

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

(ii) 任意の $k \leqq \bar{n}(E)$ に対して
$P_{\tilde{\pi}}(\sigma_k < \infty|X_0 = i_0, q) = 1$.

上記の補題5により，$q \in \mathbb{Q}^*(i_0)$ に関して初期状態 i_0 から出発したき，第 $\bar{n}(\{i_0\})$ 期以降には i_0 を含む到達可能類 $\bar{E}(q)$ を見つけることができ，次の定理を得る．

定理 5. 仮定 1のもとで，政策列 $\{\tilde{\pi}^{\tau_n}\}_{n=1}^{\infty}$ $(\tau_n \to 0 \text{ as } n \to \infty)$ は $\mathbb{Q}^*(i_0)$ に関するほとんど最適な適応型政策列である．

3.3 ニューロ・ダイナミックプログラミング

推移確率行列 $q \in \mathbb{Q}_\delta$ において確率過程 $\{X_n\}_{n=0}^{\infty}$ に対する更新関数 $\{\tilde{v}_n\}_{n=0}^{\infty}$ の確率アルゴリズムとしてニューロ・ダイナミックプログラミング[5] (Neuro-dynamic programming, Neuro-DP)における時間差分法(Temporal Difference method, TD-method)の適用について説明する．

任意の写像 $H : B(S) \to B(S)$ に関して，更新関数 $\tilde{v}_n$ を以下の方程式で各 $i \in S$ について

$$\begin{aligned} &\tilde{v}_0(i) \equiv 0, \tilde{v}_{n+1}(i) = \big(1 - \tilde{\gamma}_n(i)\big)\tilde{v}_n(i) + \\ &\tilde{\gamma}_n(i)\big(H\tilde{v}_n(i) + W_n(i) + u_n(i)\big), \ (n \geqq 0) \end{aligned} \tag{18}$$

と与える．ただし，$\tilde{\gamma}_n(i)$ は時刻 n でのステップサイズを表し，前もって与えられる $\{\gamma_n(i)\}$ によって

$$\tilde{\gamma}_n(i) = \begin{cases} \gamma_n(i), & X_n = i \text{ のとき}, \\ 0, & \text{その他} \end{cases}$$

と定義される．また，$\{W_n(i)\}$ と $\{u_n(i)\}$ はともに各状態 $i \in S$ でのランダムノイズ(random noise)を表す．

補題 6.(cf. Proposition 4.5 in [5]) 以下の条件 (i) – (v) が成り立つと仮定する．

(i) 各 $i \in S$ に対して $E[W_n(i)|\mathcal{F}_n] = 0$.

(ii) ある $A, B > 0$ が存在して
$E\big[W_n(i)^2 \mid \mathcal{F}_n\big] \leqq A + B\|\tilde{v}_n\|^2 \ (n \geqq 0, i \in S)$
が成り立つ．

(iii) H は一意の不動点 $v^* \in B(S)$ を持つ縮小写像である．

(iv) $\tilde{\gamma}_n(i) \geqq 0, \sum_{n=0}^{\infty} \tilde{\gamma}_n(i) = \infty, \sum_{n=0}^{\infty} \tilde{\gamma}_n(i)^2 < \infty \ (n \geqq 0, i \in S)$.

(v) $i \in S, n \geqq 0$ について $|u_n(i)| \leqq \theta_n(||\tilde{v}_n|| + 1)$ を満たす非負の確率変数列 $\{\theta_n\}$ が存在し，確率1で $\{\theta_n\}$ は0に収束する．

このとき，式(18)の $\tilde{v}_n$ は確率1で v^* に収束する．ただし $||\cdot||$ はsupノルム(supremum norm)であり $\mathcal{F}_n$ は $\{X_\ell(\ell \leqq n), W_\ell(\ell \leqq n-1), U_\ell(\ell \leqq n-1)\}$ によって生成される最小のσ-集合体を表す．

補題 7. 政策 $\pi = (\pi_0, \pi_1, \cdots) \in \Pi$ は状態 $i, j \in S$ と推移確率行列 $q \in \mathbb{Q}_\delta$ について

$$\pi_n\big(A^*(j|q) \mid X_0, \Delta_0, \cdots, \Delta_{n-1}, X_n = j\big) \to 1 \ (n \to \infty) \text{ with } P_\pi(\cdot|X_0 = i, q)\text{-prob. } 1$$

であるとき，π は $\mathbb{Q}_\delta$ に関して適応型最適政策である．

$q \in \mathbb{Q}_\delta$ に関する学習アルゴリズムを構成するために，更新関数 $\tilde{v}_n$ と適応型政策 $\tilde{\pi}_n$ を以下のように定める:

$$\tilde{v}_0 \equiv 0, \tilde{v}_{n+1}(i) = \big(1 - \tilde{\gamma}_n(i)\big)\tilde{v}_n(i) + \tilde{\gamma}_n(i)\big(r(i, \Delta_n) + \tilde{v}_n(X_{n+1}) - \delta \sum_{\ell \in S} \tilde{v}_n(\ell)\big) \quad (n \geqq 0) \quad (19)$$

$$\tilde{\pi}_0(a|i) > 0 \ (a \in A, i \in S),$$

$$\tilde{\pi}_{n+1}(a|i) = \begin{cases} \dfrac{\varepsilon_n(i)}{K(i) - 1}, & a \neq \tilde{a}_{n+1}(i) \text{ のとき}, \\ 1 - \varepsilon_n(i), & a = \tilde{a}_{n+1}(i) \text{ のとき}. \end{cases} \quad (n \geqq 0) \quad (20)$$

ただし，$\tilde{a}_{n+1}(i)$ は式(19)によって求められた更新関数 $\tilde{v}_{n+1}$ に関する次式のmaximizerのひとつ

$$\tilde{a}_{t+1}(i) \in \arg\max_{a \in A}\Big\{r(i, a) + \sum_{j \in S} \tilde{q}^n_{ij}(a)\tilde{v}_{n+1}(j)\Big\} \ (i \in S) \quad (21)$$

であり $K(i)$ は状態 i での選択可能な決定の個数を表す．

更新関数の収束と適応型最適政策の存在を示すために次の仮定をするとき，以下の結果が示される．

仮定 2.

(i) $\lim_{t \to \infty} \varepsilon_t(i) = 0, \sum_{t=0}^{\infty} \varepsilon_t(i) = \infty$,

(ii) すべての $i \in S$ に対して $\gamma_t(i) \geqq 0$, $\sum_{t=0}^{\infty} \gamma_t(i) = \infty, \sum_{t=0}^{\infty} \gamma_t(i)^2 < \infty$.

補題 8. 仮定2の(i)の条件を満たし，$q \in \mathbb{Q}_\delta$ $(\delta > 0)$ であるとする．このとき，

(i) $j \in S, a \in A$ に対して $\lim_{t \to \infty} N_t(j|a) = \infty$ with $P_{\tilde{\pi}}(\cdot|X_0 = i, q)$-prob. 1,

(ii) $i, j \in S, a \in A$ に対して $q^t_{ij}(a) \to q_{ij}(a)$ $(t \to \infty)$ with $P_{\tilde{\pi}}(\cdot|X_0 = i, q)$-prob. 1.

定理 6. 仮定2の条件(i), (ii)を満たし，$q \in \mathbb{Q}_\delta$ $(\delta > 0)$ であるとする．このとき，$\tilde{v}_t(i) \to h(q)(i)$ $(t \to \infty)$ with $P_{\tilde{\pi}}(\cdot|X_0 = i, q)$-prob. 1.

4. 学習アルゴリズム

この節では，これまでに示した3つのモデルのそれぞれについて各期の状態推移を観測しながら適応型政策を得るための学習アルゴリズムをまとめる．

前節で示したように到達可能行列 $\mathbb{Q}^*$ と正規到達可能行列 $\mathbb{Q}^*(i_0)$ に関する学習アルゴリズムでは割引率 $(1-\tau_k)$ に関してアルゴリズムを実行し $\tau_k \to 0$ $(k \to \infty)$ とすることで漸近的に最適となる適応型政策 $\tilde{\pi}^{\tau_k} = (\tilde{\pi}^{\tau_k}_0, \tilde{\pi}^{\tau_k}_1, \ldots)$ を得ることができる．図2, 図3にそれぞれ $\mathbb{Q}^*$ と $\mathbb{Q}^*(i_0)$ のアルゴリズムを示す．どちらの場合も仮定1の条件が満たされること，式(14)を満たす狭義単調増加関数 ϕ を用いることに注意する．

Step 1. $n = 0$とせよ．τ $(0 < \tau < 1)$を一つ選び固定せよ．$\tilde{v}_0(i) = 0$ $(i \in S)$ とせよ．$\tilde{\pi}_0 \in P(A|S)$ を $\tilde{\pi}_0(a|i) > 0$ $(a \in A, i \in S)$ となるように任意に決めよ．$q^0_{ij}(a)$ を任意に選べ．

Step 2. 現在の状態 $X_n = i$ に応じて決定 $a_i \in A(i)$ を政策 $\tilde{\pi}^\tau_n$ から選び次の期の状態 $X_{n+1} = j$ を観測せよ．そして $N_n(i, j|a), N_n(i|a)$ を計算し

$$\tilde{q}^n_{ij}(a) = \begin{cases} \dfrac{N_n(i, j|a)}{N_n(i|a)} & (N_n(i|a) > 0 \text{ のとき}), \\ q^0_{ij}(a) & (\text{その他のとき}), \end{cases}$$

とおけ．

Step 3. 各状態 $i \in S$ について $\tilde{a}_{n+1}(i)$ を

$$\tilde{a}_{n+1}(i) \in \arg\max_{a \in A}\Big\{r(i, a) + (1 - \tau)\sum_{j \in S} \tilde{q}^n_{ij}(a)\tilde{v}_n(i)\Big\}$$

となるように選べ．

Step 4. $\tilde{a}_i = \tilde{a}_{n+1}(i)$ $(i \in S)$ と表す時，次の期の政策 $\tilde{\pi}^\tau_{n+1}(i)$ $(i \in S)$ を

$$\tilde{\pi}^\tau_{n+1}(\alpha|i) = \phi(\tilde{\pi}^\tau_n(\alpha|i)) \ (\alpha \neq \tilde{a}_i)$$

$$\tilde{\pi}^\tau_{n+1}(\tilde{a}_i|i) = 1 - \sum_{\alpha \neq \tilde{a}_i} \phi(\tilde{\pi}^\tau_n(\alpha|i))$$

として更新せよ．

さらに，$\tilde{v}_{n+1} = U_\tau\{\tilde{q}^n\}\tilde{v}_n$ によって $\tilde{v}_n$ を更新せよ．

Step 5. n を $n+1$ としてStep 2へ戻れ．

図2 $\mathbb{Q}^*$ に関する学習アルゴリズム

Step 1. $n=0$とせよ. τ $(0<\tau<1)$を一つ選び固定せよ. $E_0=\{i_0\}, T_0=S-E_0, \tilde{v}_0(i)=0$ $(i\in E_0)$, $X_0=i_0$ とせよ. 各決定 $a\in A(i_0)$ に対して $\tilde{\pi}_0^\tau(a|X_0)>0$ となるように任意に決めよ.

Step 2. 決定 $\Delta_{n+1}=a_{n+1}\in A(X_n)$ を政策 $\tilde{\pi}_n(\cdot|H_n)$から選べ. 次の期の状態 $X_{n+1}=j$ を観測し,

$$E_{n+1}=\begin{cases}E_n\cup\{X_{n+1}\}, & X_{n+1}\in T_n \text{ のとき},\\ E_{n+1}=E_n, & X_n\in E_n\text{のとき}\end{cases}$$

とせよ. $i,j\in E_{n+1}, a\in A(i)$ について $N_n(i,j|a), N_n(i|a)$ を計算し,

$$\tilde{q}_{ij}^n(a)=\begin{cases}\dfrac{N_n(i,j|a)}{N_n(i,a)}, & N_{n+1}(i|a)>0 \text{ のとき}\\ q_j^0, & \text{その他}\end{cases}$$

とせよ. ただし, $q^0=(q_j^0: j\in E_{n+1})$ は E_{n+1} 上の $q_j^0>0$ $(i\in E_{n+1})$ であるような任意の確率分布である.

Step 3. 各状態 $i\in E_{n+1}$ について $\tilde{a}_{n+1}(i)$ を

$$\tilde{a}_{n+1}(i)\in\arg\max_{a\in A(i)}\{r(i,a)+(1-\tau)\sum_{j\in E_{n+1}}\tilde{q}_{ij}^n(a)\tilde{v}_n(j)\}$$

を満たすように選べ.

Step 4. $\tilde{a}_i=\tilde{a}_{n+1}(i)$ と表す時, $\tilde{\pi}_{n+1}^\tau(\alpha|i)=Prob.(\Delta_{n+1}=\alpha|H_n,\Delta_n,X_{n+1}=i)$ $(\alpha\in A(i))$ を以下のように更新せよ:

$$\tilde{\pi}_{n+1}^\tau(\alpha|i)=\phi(\tilde{\pi}_n^\tau(\alpha|i))\ (\alpha\neq\tilde{a}_i)$$
$$\tilde{\pi}_{n+1}^\tau(\tilde{a}_i|i)=1-\sum_{\alpha\neq\tilde{a}_i}\phi(\tilde{\pi}_n^\tau(\alpha|i))$$

さらに, E_{n+1} 上で $\tilde{v}_{n+1}=U_\tau\{\tilde{q}^n\}\tilde{v}_n$ によって $\tilde{v}_n$ を更新せよ.

Step 5. n を $n+1$ としてStep 2へ戻れ.

図3 $\mathbb{Q}^*(i_0)$ に関する学習アルゴリズム

最後に, 仮定 2のもとで, マイノリゼーション条件を満たす行列 $\mathbb{Q}_\delta$ に関するアルゴリズムを示す.

Step 1. $n=0, \tilde{v}_0\equiv 0$ とし $\tilde{\pi}_0\in P(A|S)$ を $\tilde{\pi}_0(a|i)>0$ $(a\in A, i\in S)$ となるように任意に一つ決めよ.

Step 2. $\Delta_n=a_n$ を政策 $\tilde{\pi}_n$ に従って一つ選べ. $X_n=i$ と a_n から次の期の状態 $X_{n+1}=j$ を観測せよ. $n+1$ 期において $\tilde{v}_{n+1}\in B(S)$ の値を次の時間差分方程式によって更新せよ:

状態 $i\in S$ に対して,

$$\tilde{v}_{n+1}(i)=\big(1-\tilde{\gamma}_n(i)\big)\tilde{v}_n(i)+\tilde{\gamma}_n(i)\big(r(i,\Delta_t)+\tilde{v}_n(X_{n+1})-\delta\sum_{\ell\in S}\tilde{v}_n(\ell)\big)$$

ただし, ステップサイズ $\tilde{\gamma}_n$ は事前に与えられた $\{\gamma_n(i)\}$ により

$$\tilde{\gamma}_t(i)=\begin{cases}\gamma_n(i), & X_n=i \text{ のとき},\\ 0, & \text{その他}\end{cases}$$

とする.

Step 3. 各状態 $i\in S$ に対して

$$\tilde{a}_{n+1}(i)\in\arg\max_{a\in A}\{r(i,a)+\sum_{j\in S}\tilde{q}_{ij}^n(a)\tilde{v}_{n+1}(j)\}$$

を選び, 政策 $\tilde{\pi}_{n+1}$ を次のように決めよ:

$$\tilde{\pi}_{n+1}(a|i)=\begin{cases}\frac{\varepsilon_n(i)}{K(i)-1}, & a\neq\tilde{a}_{n+1}(i) \text{ のとき},\\ 1-\varepsilon_n(i), & a=\tilde{a}_{n+1}(i) \text{ のとき}.\end{cases}$$

ただし $K(i)$ は状態 $i\in S$ での決定の個数を表す.

Step 4. $n=n+1$ として Step 2へ戻れ.

図4 $\mathbb{Q}_\delta$ に関する学習アルゴリズム

5. おわりに

本稿では, 不確実性の下でのマルコフ決定過程における3つの適応型アルゴリズムについて示した. 未知の推移確率行列のモデル構造はそれぞれ異なるが, いずれも動的システムでの意思決定と状態観測に基づき探索(exploration)と知識利用(exploitation)のトレードオフをうまく行いながら次の期の政策改善を逐次行う学習アルゴリズムであり, その手順も簡明である. 理論の詳細や数値例については[11, 12, 27]を参照されたい.

謝辞 本稿をまとめるにあたり, 伊喜哲一郎氏, 蔵野正美氏, 安田正實氏の各氏との共同研究の内容をもとにさせて頂きました. ここに感謝の意を表します.

参考文献

[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] J. Bather. Optimal decision procedures for finite Markov chains. II. Communicating sys-

tems. *Advances in Appl. Probability*, 5:521–540, 1973.

[3] R. Bellman. *Dynamic programming*. Princeton, 1957.

[4] R. Bellman. A Markovian decision process. *J. Math. Mech.*, 6:679–684, 1957.

[5] D. P. Bertsekas and J. H. Tsitsiklis. *Neuro-Dynamic Programming*. Athena, 1996.

[6] 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):447–469, 2000.

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

[8] O. Hernández-Lerma. *Adaptive Markov control processes*. Springer, 1989.

[9] R. A. Howard. *Dynamic programming and Markov processes*. M.I.T., 1960.

[10] T. Iki, M. Horiguchi, and M. Kurano. A structured pattern matrix algorithm for multichain Markov decision processes. *Math. Methods Oper. Res.*, 66:545–555, 2007.

[11] T. Iki, M. Horiguchi, M. Yasuda, and M. Kurano. A learning algorithm for communicating Markov decision processes with unknown transition matrices. *Bull. Inform. and Cybernet.*, 39:11–24, 2007.

[12] 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 Int. conf. Proc. MDAI 2007 (CD-ROM Proceedings)*, pages 112–122, 2007.

[13] V. R. Konda and V. S. Borkar. Actor-critic-type learning algorithms for Markov decision processes. *SIAM J. Control Optim.*, 38(1):94–123, 1999.

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

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

[16] S. Lakshmivarahan. *Learning algorithms*. Springer, 1981.

[17] P. Mandl. Estimation and control in Markov chains. *Adv. in Appl. Prob.*, 6:40–60, 1974.

[18] J. J. Martin. *Bayesian decision problems and Markov chains*. John Wiley & Sons Inc., 1967.

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

[20] G. E. Monahan. A survey of partially observable Markov decision processes: theory, models, and algorithms. *Management Sci.*, 28(1):1–16, 1982.

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

[22] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons Inc., 1994.

[23] S. M. Ross. *Applied probability models with optimization applications*. Holden-Day, 1970.

[24] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

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

[26] 伊喜哲一郎, 堀口正之. A modified pattern matrix algorithm for multichain MDPs. RIMS講究録1504「情報決定過程論の展開」, pages 73–86, 2006.

[27] 伊喜哲一郎, 堀口正之, 蔵野正美, 安田正實. A pattern-matrix learning algorithm for adaptive MDPs:the regularly communicating case. RIMS講究録1589「不確実な状況における意思決定の理論と応用」, pages 110–119, 2008.

[28] 北川敏男編, 小河原正巳, 坂本武司著. マルコフ過程. 共立出版, 1967.

[29] 堀口正之, 蔵野正美, 安田正實. マルコフ決定過程におけるTD法による学習アルゴリズムについて. RIMS講究録1559「最適化問題における確率モデルの展開と応用」, pages 34–49, 2006.