数理解析研究所講究録 1306 巻 2003 年 101-108

マルコフ決定過程に対する Q-learning について

徳島大学総合科学部　大橋　守(Mamoru Ohashi)
Faculty of Integrated Arts and Sciences
The University of Tokushima

1　はじめに

マルコフ決定過程は広い応用範囲を持っている．最適な方策は，マルコフ決定過程に対する DP 方程式を解くことにより求めることができる．しかし，事前に状態推移確率と推移に関係した費用を推定する必要がある．一般に，これらの値を正確に知ることはできない．ここでは，状態推移の実現値とその推移に関係した費用をもとに DP 方程式の最適費用と値関数および最適方策を求める問題を取り扱う．

Abounadi, Bertsekas and Borkar [1] は平均費用を最小にする問題について議論している．彼らは，Q-learning と呼ばれる学習則を一部改良して，DP 方程式の最適費用と値関数にほとんど確実に収束する学習則を示した．その方法は ODE 法[2,3]とマルチンゲールの収束定理を用いている．また，Q-learning の数値計算については Mahadevan[4]が現実的な例に対して詳しく議論している．本稿では，彼らの学習則がほとんど確実に収束し，その期待値が DP 方程式の最適費用と値関数に一致することを示す．さらに，マルコフ決定過程を用いてモデル化した保全問題に対してこの学習則を適用し，最適費用と最適保全方策を求める．この例を通して彼らの学習則の収束性について考察する．

平均費用を最小にする保全問題の数値例では，DP 方程式の解である最適費用と値関数の関数値の近傍までは比較的早く収束するが，その後は揺れが長く続く．学習則の初期値として DP 方程式の解である最適費用とそのときの値関数を用いた場合も同様の結果となる．彼らの学習則の収束は非常に遅いと考えられる．しかし，最適方策への収束は比較的早い，

2　マルコフ決定過程

有限状態空間が $S=\{1,2,\cdots,d\}$，有限決定空間が $A=\{a_1,a_2,\cdots,a_r\}$ のマルコフ決定過程 $\{X_n, n\geq 0\}$ を考える．決定 $a\in A$ のもとで，状態 $i\in S$ から状態 $j\in S$ に推移する確率を $p(i,a,j)$，この推移に関係した費用を $g(i,a,j)$ とする．このとき，次の仮定 A1 のもとで平均費用

$$(1)\qquad \limsup_{n\to\infty}\frac{1}{n}\sum_{m=0}^{n-1}E[g(X_m,Z_m,X_{m+1})],$$

を最小にする決定の列 $\{Z_n\}$ を決める定常方策 $Z_n=v(X_n)$ と最小値 β を求める．

マルコフ決定過程 $\{X_n, n \geq 0\}$ *はすべての定常方策* $v: S \to A$ *に対して既約である.*

Ross[5]の6章より (1) 式の最小値である最適費用 β と値関数 $V(i): S \to R$ は次のDP方程式を満たす.

$$(2) \qquad V(i) = \min_{a \in A} \left[\sum_{j \in S} p(i,a,j)(g(i,a,j) + V(j)) - \beta \right], \quad i \in S.$$

いま, 上式の右辺の括弧の中を $Q(i,a): S \times A \to R$ とおくと, 最適方策は $v(i) \in \mathrm{Arg}\min(Q(i,\cdot))$ と表わすことができる. この $Q(\cdot,\cdot)$ をQ-factorといい, 定数の違いを除いて唯一つに定まり, $V(i) = \min_{a \in A} Q(i,a)$ となる. よって,

$$(3) \qquad Q(i,a) = \sum_{j \in S} p(i,a,j)(g(i,a,j) + \min_{b \in A} Q(j,b)) - \beta, \quad j \in S, a \in A,$$

と書くこともできる. Q-learningと呼ばれる学習則は状態推移確率 $p(i,a,j)$ が未知あるとき, 確率分布 $p(i,a,\cdot)$, $i \in S, a \in A$ に従い, 互いに独立な確率変数 $\xi_{i,a}^n$ の実現値を用いてQ-factorを求める方法のことである.

3 Q-learning

DP方程式 (2) の最適費用と値関数を求める代表的なRVI(relative value iteration) は

$$(4) \qquad V^{n+1}(i) = \min_{a \in A} \left[\sum_{j \in S} p(i,a,j)\left(g(i,a,j) + V^n(j)\right) - V^n(i_0) \right], \quad i \in S,$$

$$V^n(i_0) \to \beta, \quad n \to \infty,$$

である. ただし, $i_0 \in S$ は任意とする. このとき, $V^n(i)$ は $V(i_0) = \beta$ であるDP方程式 (2) の解 $V(i)$ に収束する. 同様に, (3) 式のQ-factorは

$$(5) \qquad Q^{n+1}(i,a) = \sum_{j \in S} p(i,a,j)(g(i,a,j) + \min_{b \in A} Q^n(j,b)) - Q^n(i_0,a_0), \quad i \in S, a \in A,$$

$$Q^n(i,a) \to Q(i,a), Q^n(i_0,a_0) \to \beta, \quad n \to \infty,$$

を用いて計算できる. ただし, $i_0 \in S, a_0 \in A$ は任意とする.

これらの (4), (5) を用いて最適費用 β と値関数 $V(i)$ およびQ-factor $Q(i,a)$ を計算することができる. しかし, 状態推移確率 $p(i,a,j)$ と推移に関係した費用 $g(i,a,j)$ が既知でな

ければならない．一般に，これらの値を推定するのは困難な場合が多い．Abounadi, Bertsekas and Borkar[1]は Q-learning の学習則を一部改良して，以下の仮定 A2 のもとで，状態推移の実現値を用いた次の学習則を示した．

$$Q^{n+1}(i,a)=Q^n(i,a)+\gamma(n)\{g(i,a,\xi_{ia}^n)+\min_b Q^n(\xi_{ia}^n,b) \\ -Q^n(i_0,a_0)-Q^n(i,a)\},\quad i\in S,a\in A. \tag{6}$$

ただし，$i_0\in S, a_0\in A$ は任意で，$\xi_{i,a}^n$ は互いに独立で確率分布 $p(i,a,\cdot)$, $i\in S,a\in A$ に従う確率変数の実現値とする．

仮定 A2 (tapering stepsize)

ステップサイズ $\gamma(n)>0$ は

$$\sum_k \gamma(k)=\infty,\quad \sum_k \gamma^2(k)<\infty$$

とする.

学習則 (6) 式は$T:R^{d\times r}\to R^{d\times r}$を

$$(TQ^n)(i,a)=\sum_{j\in S} p(i,a,j)(g(i,a,j)+\min_{b\in A} Q^n(j,b)),$$

とおくと，マルチンゲール

$$M^{n+1}(i,a)=g(i,a,\xi_{ia}^n)+\min_{b\in A} Q^n(\xi_{ia}^n,b)-(TQ^n)(i,a),$$

を用いて

$$Q^{n+1}=Q^n+\gamma(n)(T(Q^n)-Q^n(i_0,a_0)e-Q^n+M^{n+1}), \tag{7}$$

と書き換えることができる．ただし，$Q^n,M^n\in R^{d\times r}$,

$$e=\begin{pmatrix}1\\1\\\vdots\\1\end{pmatrix}\in R^{d\times r},$$

とする．いま，$\Im_n=\sigma(Q^m,M^m,m\le n)$, $n\ge 0$ とおくと，マルチンゲール$\{M^n,n\ge 1\}$ は

$$E[M^{n+1}\mid \Im_n]=0,\quad E\left[\left\|M^{n+1}\right\|^2\mid \Im_n\right]\le K(1+\left\|Q^n\right\|^2) \tag{8}$$

となる．さらに，ODE 法[2,3] を用いるために次の微分方程式

$$\dot{Q}(t)=T(Q(t))-Q(t)(i_0,a_0)e-Q(t), \tag{9}$$

を考える．Abounadi, Bertsekas and Borkar [1]は（7），（8），（9）を用いて次の結果を得た．

補題 1（[1] の Lemma 3.8）

$\{Q^n\}$ はほとんど確実に有界である．

補題 2（[1] の Theorem 3.4）

Q は微分方程式（9）の大域的漸近安定である．

さらに，彼らはこれらの補題 1，2 より次の結果を示した．

定理 3（[1] の Theorem 3.5）

Q^n はほとんど確実に Q に収束する．

この定理より Q^n は Q に概収束することからマルコフ決定過程の状態推移の実現値とその推移に関係した費用の実現値より，(6) 式を用いて最適費用と値関数を求めることができる．しかし，すべての $i \in S, a \in A$ に対して確率変数 $\xi_{i,a}^n$ の実現値を求める必要がある．

確率論ではマルチンゲール $\{M^n, n \geq 1\}$ の収束に関して次の定理が良く知られている．

定理 4（J. ヌブ[6] の命題 4.6.1）

マルチンゲール $\{M^n, n \geq 1\}$ に対して

(1) *級数 $\sum_{n\geq 1} E\|M^n\|^2$ が収束すれば，極限 $\lim_{n\to\infty} \sum_{m=1}^{n} M^n$ が a.s. の意味と L_2 の意味で存在する．*

(2) *∞ に発散する実数の任意の非減少列 $\{u_n, n \geq 1\}$ に対して，級数 $\sum_{n\geq 1} u_n^{-2} E\|M^n\|^2$ が収束すれば，$\lim_{n\to\infty} u_n^{-1} \sum_{m=1}^{n} M^n = 0$ a.s. となる．*

この定理と補題 1，2 より定理 3 より弱い次の定理を得ることができる．

定理 5

Q^n はほとんど確実に収束し，$E\left[\lim_{n\to\infty} Q^n\right] = Q$ となる．

[証明]（7）式は次のように書き換えることができる．

(10) $$Q^n = Q^0 + \sum_{m=0}^{n} \gamma(m)(T(Q^m) - Q^m(i_0,a_0)e - Q^m) + \sum_{m=0}^{n} \gamma(m)M^{m+1}.$$

(8) 式より上式の第3項は

$$E\left[\gamma(n)M^{n+1} \mid \Im_n\right] = \gamma(n)E\left[M^{n+1} \mid \Im_n\right] = 0, \quad n \geq 0.$$

また，仮定A2と補題1より

$$\sum_{n\geq 0} E\left[\left\|\gamma(n)M^{n+1}\right\|^2 \mid \Im_n\right] = \sum_{n\geq 0} |\gamma(n)|^2 E\left[\left\|M^{n+1}\right\|^2 \mid \Im_n\right]$$
$$\leq K(1+\left\|Q^n\right\|^2)\sum_{n\geq 0}|\gamma(n)|^2 < \infty$$

となる．よって，定理4より，$\sum_{m=1}^{n}\gamma(m)M^{m+1}$ はほとんど確実に収束する．さらに，$\sum_{m=1}^{n}\gamma(m)M^{m+1}$ は同程度積分可能であるから $E\left[\lim_{n\to\infty}\sum_{m=1}^{n}\gamma(m)M^{m+1}\right]=0$ となる．

補題1より Q^n は有界で，かつ，補題2より (10) 式の第1項と第2項はDP方程式の解 Q に収束する．したがって，Q^n はほとんど確実に収束し，$E\left[\lim_{n\to\infty} Q^n\right] = Q$ となる．

[注] 定理4より $\gamma(n)\sum_{m=1}^{n}M^{m+1}$ はほとんど確実に 0 に収束するが，$\sum_{m=1}^{n}\gamma(m)M^{m+1}$ はほとんど確実に収束するが，0 に収束するとは限らない．Abounadi, Bertsekas and Borkar [1]の定理3は Q^n がほとんど確実に Q 収束することを示している．

4 数値例

この節では，平均費用を最小にする5状態，2決定の保全問題をマルコフ決定過程でモデル化し，(7) の学習則を用いて最適費用と最適保全方策を求める．数値計算には表計算ソフトを利用し，乱数を用いて確率変数 $\xi_{i,a}^n$ の実現値をシミュレーションした．

状態空間を $S=\{1,2,\cdots,5\}$，決定空間を $A=\{1,2\}$ とする．また，状態推移確率 $p(i,a,j)$ を

$$P(\cdot,1,\cdot)=\begin{pmatrix} 0.2 & 0.3 & 0.3 & 0.1 & 0.1 \\ 0 & 0.3 & 0.3 & 0.2 & 0.2 \\ 0 & 0 & 0.4 & 0.3 & 0.3 \\ 0 & 0 & 0 & 0.5 & 0.5 \\ 0.1 & 0 & 0 & 0 & 0.9 \end{pmatrix}, \qquad P(\cdot,2,\cdot)=\begin{pmatrix} 0.3 & 0.4 & 0.2 & 0.1 & 0 \\ 0 & 0.4 & 0.3 & 0.3 & 0 \\ 0 & 0 & 0.5 & 0.4 & 0.1 \\ 0 & 0 & 0 & 0.6 & 0.4 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix},$$

推移に関係した費用 $g(i,a,j)$ を

$$g(\cdot,1,\cdot)=\begin{pmatrix}1&1&1&1&10\\1&1&1&1&10\\1&1&1&1&10\\1&1&1&1&10\\20&1&1&1&10\end{pmatrix},\qquad g(\cdot,2,\cdot)=\begin{pmatrix}2&3&3&3&10\\2&3&3&3&10\\2&2&3&3&10\\2&2&2&2&10\\7&2&2&2&10\end{pmatrix}$$

とする．このとき，(5) 式により状態推移確率 $p(i,a,j)$ と推移に関係した費用 $g(i,a,j)$ が既知であるとき，最適費用と最適保全方策は

$$\beta=4.195652,$$
$$\{v(i)\}=(1\quad 1\quad 1\quad 2\quad 2)$$

となる．また，$i_0=1, a_0=1$ のとき Q-factor の値関数は

$$\{Q(i,a)\}=\begin{pmatrix}4.195652&5.907609&7.429348&9.559783&13.52391\\4.563043&6.249457&7.723370&9.510870&7\end{pmatrix}$$

となる．

次に，状態推移確率 $p(i,a,j)$ に従う確率変数の実現値を表計算ソフトの乱数を用いて発生させ，学習則 (6) に従って最適費用と Q-factor の値関数を計算した．表計算ソフトの乱数には問題も多いが収束の様子を見るには十分であると考えられる．次の図 1 は Q-factor の初期値として $Q^0(i,a)=0,\ i\in S, a\in A, i_0=1, a_0=1$ を用いたとき，最適費用 β に収束する $Q^n(i_0,a_0)$ のグラフである．ただし，$\gamma(n)=\dfrac{1}{1+n}$ で，右のグラフはほぼ収束した状態での $Q^n(i_0,a_0)$ の様子を拡大したものである．最適費用 $\beta=4.195652$ の近傍で長く変動している．次に，初期値として DP 方程式の解を用いたとき，$Q^n(i_0,a_0)$ の収束の様子を図 2 に示す．同じスケールで $\gamma(n)$

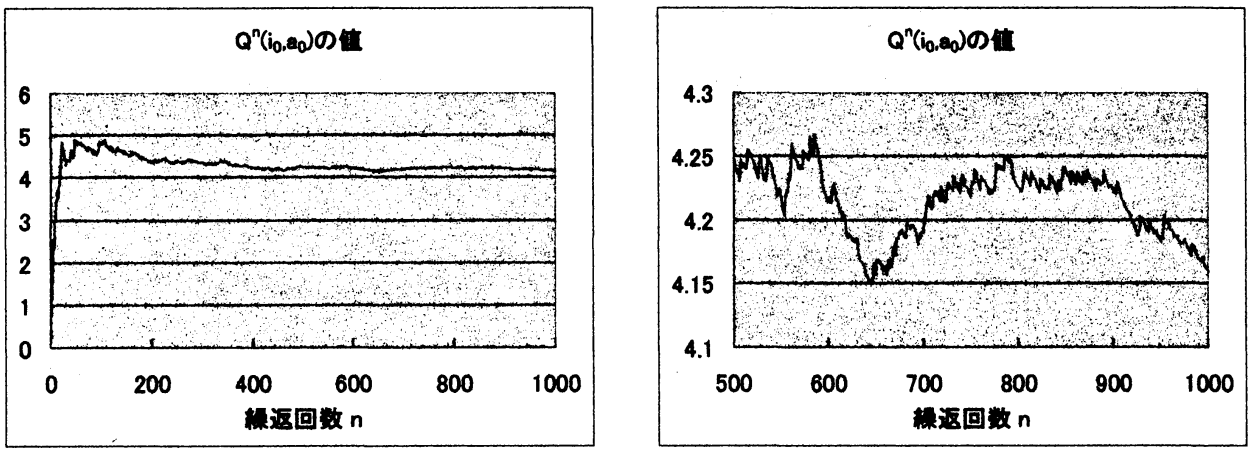
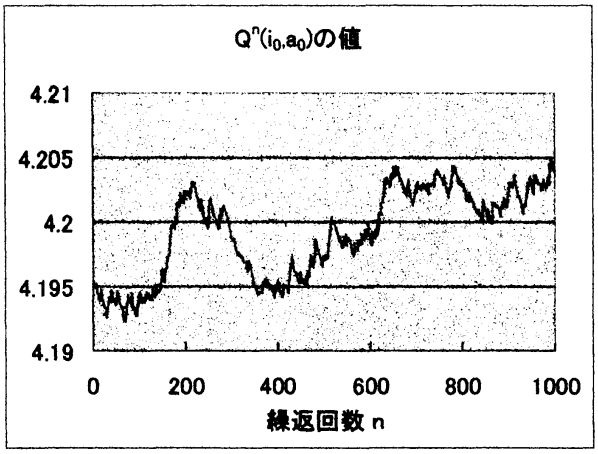
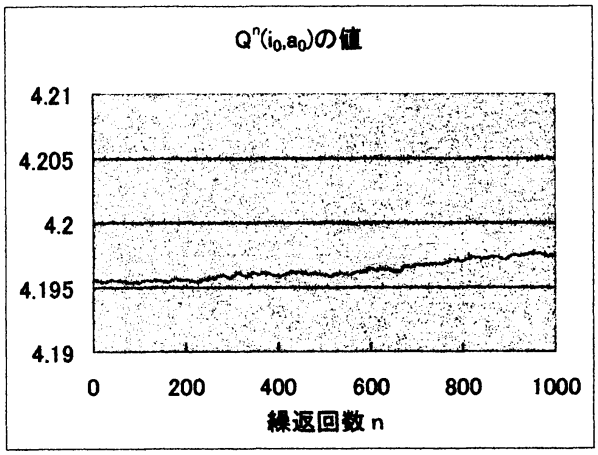


図 1 $Q^0(i,a)=0,\ \gamma(n)=\dfrac{1}{1+n}$ のときの収束





$$\gamma(n)=\frac{1}{10000+n}$$

$$\gamma(n)=\frac{1}{100000+n}$$

図 2 $Q^0(i,a)=Q(i,a)$, $Q^0(i_0,a_0)=\beta$ のときの収束

を変え，左側が $\gamma(n)=\dfrac{1}{10000+n}$，右側が $\gamma(n)=\dfrac{1}{100000+n}$ の場合のグラフで，$\gamma(n)$ の選び方によって変動の範囲が変わる．しかし，図 3 は右側のグラフを縦軸方向に拡大したものである．これらの図より $Q^n(i_0,a_0)$ は最適費用 $\beta=4.195652$ の近傍で変動し，$\gamma(n)$ の選び方によって変動の範囲が変わる．図 4, 5 は最適保全方策の変化を示したものである．図 4 に Q-factor の初期値として $Q^0(i,a)=0$，

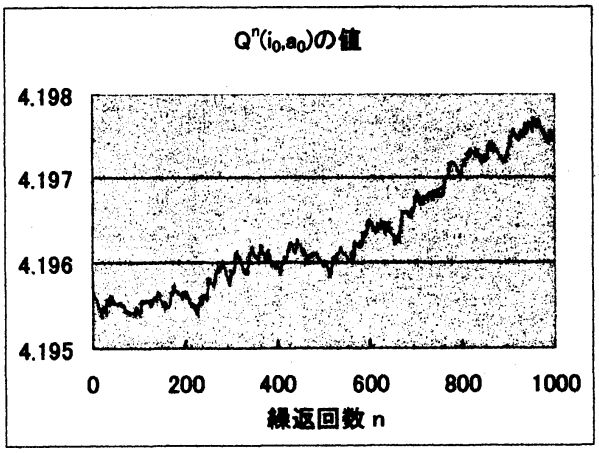


図 3 上のグラフの拡大図

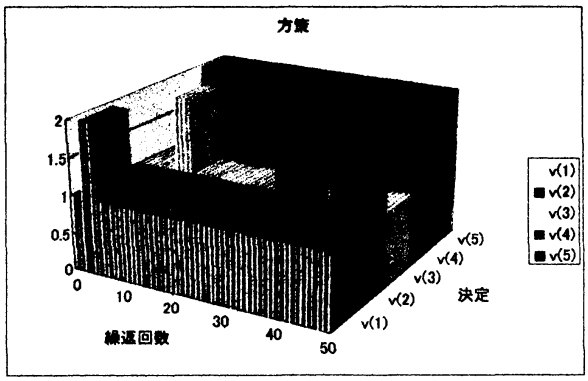


図 5 $Q^0(i,a)=0$のときの最適保全方策

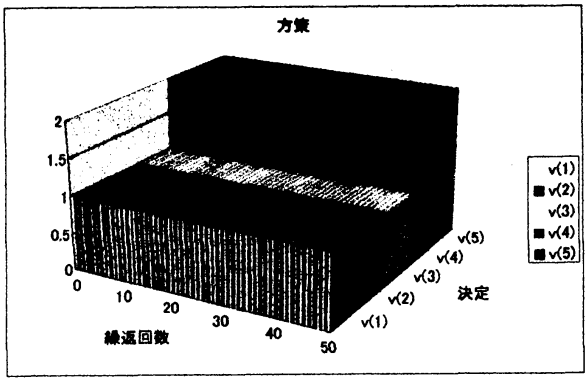


図 6$Q^0(i,a)=Q(i,a)$のときの最適保全方策

$\gamma(n)=\frac{1}{1+n}$ を用いたとき，最初の 50 回の変化を示し，図 5 には初期値として DP 方程式の解と $\gamma(n)=\frac{1}{10000+n}$ を用いたときの 50 回の変化を示した．数値計算の結果，最適保全方策は比較的早く収束し，最適値の近傍での変動が少ない．

5 終わりに

状態推移の実現値とその推移に関係した費用の実現値を用いた学習則 (7) による概収束は非常に遅いことが分かる．その主な理由は (10) 式第 3 項のマルチンゲールの収束の悪さに関係している．Q が微分方程式 (9) の大域的漸近安定であることから Q^n は Q に収束するが，第 3 項のマルチンゲールの収束は $\gamma(n)$ に依存している．しかし，定理 5 より Q^n はほとんど確実に収束し，その期待値が Q であることから適当なところで打ち切って，その平均値を用いることができると考えられる．また，最適方策は比較的早く収束することから最適方策を求める方法としては有効である．

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 (2001), pp.681-698.

[2] V. S. Borkar and S. P. Meyn, The ODE method for convergence of stochastic approximation and reinforcement learning, SIAM J. Control Optim., 38 (2000), pp.447-469.

[3] V. R. Konda and V. S. Borkar, Actor-critic-type learning algorithms for Markov decision processes, SIAM J. Control Optim., 38 (2000), pp.94-123.

[4] S. Mahadevan, Average reward reinforcement learning: Foundations, algorithms for Markov decision processes, SIAM J. Control Optim., 38 (2000), pp.94-123.

[5] S. M. Ross, Applied Probability Models with Optimization Applications, Holden-Day, 1969.

[6] J. ヌブ著，鶴見，大石，川尻共訳，確率論，共立出版，1974.