Vol. 39 No. 4 情報処理学会論文誌 Apr. 1998

# 学習期間と制御期間に分割された強化学習問題における最適アルゴリズムの提案

前田康成[†,☆] 浮田善文[†]
松嶋敏泰[†] 平澤茂一[†]

本研究では，遷移確率行列が未知であるようなマルコフ決定過程によってモデル化されている，学習期間と制御期間に分割された強化学習問題における，最適アルゴリズムの提案を行っている．従来研究では，真の遷移確率行列を同定できれば制御期間の収益を最大化できるため，学習期間の目的を単に未知の遷移確率行列の推定としているが，有限の学習期間のもとでは推定誤差があるため，収益最大化の厳密な保証はない．そこで本研究では，有限の学習期間と有限の制御期間の強化学習問題において，制御期間の収益をベイズ基準のもとで最大化する基本最適アルゴリズムを提案する．しかし，基本最適アルゴリズムの計算量が指数オーダーのため，さらにその改良を行い，改良最適アルゴリズムを提案する．改良最適アルゴリズムは基本最適アルゴリズム同様に収益をベイズ基準のもとで最大化することができ，かつその計算量は多項式オーダーに軽減されている．

# The Optimal Algorithms for the Reinforcement Learning Problem Separated into a Learning Period and a Control Period

YASUNARI MAEDA,[†,☆] YOSHIHUMI UKITA,[†] TOSHIYASU MATSUSHIMA[†] and SHIGEICHI HIRASAWA[†]

In this paper, new algorithms are proposed based on statistical decision theory in the field of Markov decision processes under the condition that a transition probability matrix is unknown. In previous researches on RL (reinforcement learning), learning is based on only the estimation of an unknown transition probability matrix and the maximum reward is not received in a finite period, though their purpose is to maximize a reward. In our algorithms it is possible to maximize the reward within a finite period with respect to Bayes criterion. Moreover, we propose some techniques to reduce the computational complexity of our algorithm from exponential order to polynomial order

## 1. 序　論

最近，人工知能の学習の分野において広く強化学習が研究されている．強化学習は何らかの未知情報を含むような制御問題において，未知情報について学習しながら，収益を最大化するように制御を行う問題である．つまり，強化学習は学習と制御の混合問題になっている．

強化学習における従来研究のほとんどにおいて，そのモデルはマルコフ決定過程（MDP, Markov Decision Processes）[6),8),14)] で表現されてきた．マルコフ決定過程ですべての情報が既知の場合は，単に制御の問題である．制御期間が有限の固定期間の場合には，動的計画法（DP, Dynamic Programming）[3),6),8),14)] の問題を解くことによって収益を最大化する最適政策が求められる．制御期間が無限の場合には，Policy Iteration Algorithm（PIA）[6),8),14)] によって収益を最大化する最適政策が求められる．

しかし，強化学習では未知情報を含むマルコフ決定過程，すなわち，学習と制御の混合問題が扱われる．従来研究の中には，学習に重きを置いた研究に Q-Learning[15)] がある．Q-Learning では制御期間が無限の問題に対して，その出力がすべての情報が既知の場合の最適政策（以下，真の最適政策と呼ぶ）に漸近的に収束することが保証されている．一方，制御に重きを置いた研究に Martin によって提案されたアル

† 早稲田大学理工学部経営システム工学科
Department of Industrial and Management Systems Engineering, School of Science and Engineering, Waseda University
☆ 現在，NTT 情報通信研究所
Presently with NTT Information and Communication Systems Laboratories

ゴリズム[7),12)]がある．Martinのアルゴリズムでは制御期間が有限の固定期間の問題に対して，ベイズ決定理論[3),5),13)]に基づくDPの問題を解くことによって，制御期間の収益がベイズ基準のもとで最大化される．

本研究では，学習期間と制御期間に分割して強化学習の問題を扱うことによって，強化学習問題を学習と制御に明確に分けて解くことを考える．このような問題を扱った従来研究には宮崎ら[9)~11)]やBartoら[1)]の研究がある．これらの従来研究では，無限期間の制御期間の問題に対して，学習期間にはマルコフ決定過程の遷移確率行列を支配する未知パラメータの最尤推定が行われ，制御期間には学習期間で求めたパラメータの推定値を真のものと仮定してPIAで求めた政策を用いて制御が行われる．これらの研究では，学習期間の長さが無限であれば真の最適政策への収束が保証されている．しかし，有限の学習期間のもとでは推定誤差があるため，収益最大化の厳密な保証はない．

そこで，本研究では有限で固定の学習期間と有限で固定の制御期間の問題に対して，制御期間の収益に何らかの保証を与えることが目的である．本研究では，Martinのアルゴリズムと同様にベイズ決定理論に基づくDPの問題を解くことによって，制御期間の収益をベイズ基準のもとで最大化することを考える．なお，学習期間の収益は無視する．よって，学習期間における目的は，制御期間における収益を最大化するための情報を収集することである．

まず，2章で強化学習の多くの研究におけるモデルであり，かつ，本研究のモデルでもあるマルコフ決定過程について述べる．3章で従来研究について述べる．4章で本研究で扱う問題の概要を述べるとともに定式化を行う．5章では，5.1節でMartinのアルゴリズムと同様のDPの問題を解くことによって最適解が求められる基本最適アルゴリズムを提案する．しかし，その基本最適アルゴリズムの計算量が指数オーダーであるため，5.2節で基本最適アルゴリズムを改良し，改良最適アルゴリズムを提案する．改良最適アルゴリズムでは基本最適アルゴリズムと同様に最適解が求められ，かつ，その計算量は多項式オーダーに軽減されている．最後に6章でまとめを行う．

## 2. マルコフ決定過程

強化学習における従来研究のほとんどにおいて，そのモデルはマルコフ決定過程[6),8),14)]で表現されてきた．マルコフ決定過程は有限の状態，有限の行動，遷移確率行列，そして遷移に伴う収益を示す利得関数によって定義される．マルコフ決定過程における目的は，行動を選択し，状態が遷移し，その状態遷移に伴う収益を得るという一連のプロセスを繰り返すことによって得られる総収益を最大化することである．選択するべき行動を決定する関数を政策と呼び，総収益を最大化するような政策を最適政策と呼ぶ．

マルコフ決定過程には，割引問題と非割引問題という2種類の問題がある．割引問題における目的は，次式で示される期待割引総収益の最大化である．

$$\nu = E\left(\sum_{t=0}^{\infty}\beta^t y_t\right), \tag{1}$$

ただし，$\beta(0<\beta<1)$は割引率，$y_t$は$t$期の収益である．これは総収益の値を有限の値でおさえるために，毎期の収益が割引率によって割り引かれている．

非割引問題における目的は次式で示される期待平均収益の最大化である．これは，有限の値でおさえるために，総収益が期間の長さで割られている．

$$h = E\left(\lim_{T\to\infty}\left(\sum_{t=0}^{T} y_t \Big/ T+1\right)\right). \tag{2}$$

本研究でも，マルコフ決定過程を基本モデルとし，割引問題を扱うことにする．一般的には，強化学習においては遷移確率行列と利得関数が未知な問題設定がよく扱われている．本研究では，遷移確率行列のみが未知な場合の強化学習を扱うことにする．

## 3. 従来研究

本章では，本研究と関連の深いと思われるMartinの従来研究について述べる．

有限で固定の制御期間のみからなる強化学習問題を扱った研究として，Martinによって提案されたアルゴリズム[7),12)]がある．Martinのアルゴリズムではベイズ決定理論[3),5),13)]に基づくDPの問題を解くことによって，制御期間の期待割引総収益がベイズ基準のもとで最大化されている．まず，いくつかの記号の定義を行う．

遷移確率行列を支配する連続パラメータを$\theta$と表し，そのパラメータ集合を$\Theta$，真のパラメータを$\theta^*$，$\theta^*\in\Theta$と表す．有限の状態集合を$S$，$s_i\in S$，有限の行動集合を$A$，$a_k\in A$，状態$s_i$において行動$a_k$が選択されたもとで状態$s_j$に遷移したときに得られる収益を示す利得関数を$r(s_i,a_k,s_j)$と表す．状態$s_i$において行動$a_k$が選択されたもとで，状態$s_j$に遷移する確率を示す，パラメータ$\theta$によって支配される$|S||A|\times|S|$の遷移確率行列（$|\cdot|$は集合の濃度を示す）の要素を$p(s_j|s_i,a_k,\theta)$と表す．制御期間の

長さを $T$，$t$ 期における状態を $x_t$，制御期間の初期状態を $x_0$，$t$ 期において選択された行動を $z_t$ と表し，状態 $x_0$ において行動 $z_0$ が選択され，状態 $x_1$ へ遷移し，状態 $x_1$ において行動 $z_1$ が選択され，という一連の遷移の履歴を示す遷移系列を $x_0z_0x_1z_1\cdots x_t$ と表す．パラメータ $\theta$ の事前確率密度関数を $p(\theta)$，$x_0z_0x_1z_1\cdots x_t$ という遷移をしたもとでの事後確率密度関数を $p(\theta|x_0z_0x_1z_1\cdots x_t)$ と表す．

次に，次式によって $x_0z_0\cdots x_{t-1}$ という遷移をしたもとで，状態 $x_{t-1}$ において行動 $a$ を選択して状態 $x_t$ へ遷移する確率の，遷移確率行列を支配するパラメータの事後確率密度関数による加重平均を定義する．

$$\bar{p}(x_t|x_{t-1},a,x_0z_0\cdots x_{t-1}) = \int_\Theta p(\theta|x_0z_0\cdots x_{t-1})p(x_t|x_{t-1},a,\theta)d\theta. \tag{3}$$

Martin のアルゴリズムでは毎期ごとに，その期以後の期待割引総収益を最大化する行動選択が次式によって決定される．

$$z_t(x_0z_0\cdots x_t) = \arg\max_a \sum_{x_{t+1}} \bar{p}(x_{t+1}|x_t,a,x_0z_0\cdots x_t) (r(x_t,a,x_{t+1})+\beta u'(x_0z_0\cdots x_tax_{t+1})), \tag{4}$$

ただし，$z_t(x_0z_0\cdots x_t)$ は $x_0z_0\cdots x_t$ という遷移をしたもとで選択するべき行動を示し，

$$u'(x_0z_0\cdots x_t) = \max_a \sum_{x_{t+1}} \bar{p}(x_{t+1}|x_t,a,x_0z_0\cdots x_t) (r(x_t,a,x_{t+1})+\beta u'(x_0z_0\cdots x_tax_{t+1})), \tag{5}$$

かつ

$$u'(x_0z_0\cdots x_T)=0. \tag{6}$$

本研究では有限で固定の学習期間と有限で固定の制御期間からなる強化学習問題を扱い，Martin のアルゴリズムと同様にベイズ決定理論に基づく DP の問題を解くことによって，学習期間には制御期間の期待割引総収益をベイズ基準のもとで最大化するための行動選択を行い，制御期間の期待割引総収益をベイズ基準のもとで最大化する．

## 4. 本研究で扱うパラダイム

### 4.1 記号の定義

以下に，本研究で新たに導入する記号と，従来研究の Martin のアルゴリズムにおける記号と意味が異なる記号について定義を行う．その他の記号については Martin のアルゴリズムと同様である．

学習期間の初期状態を $x_0$，制御期間の初期状態（学習期間の最後には $x_N$ に到達するが，制御期間は $x_{N'}$ から始めることにする）を $x_{N'}$ と表す．学習期間と制御期間の各期において選択するべき行動を決定する関数である $N+T$ 期間（学習期間と制御期間）の政策を $\pi(\cdot,\cdots)$ と表し，$x_0z_0x_1z_1\cdots x_t$ という遷移をしたもとで，政策 $\pi(\cdot,\cdots)$ によって定まる $t$ 期に選択するべき行動を $\pi(t,x_0z_0x_1z_1\cdots x_t)$ と表す．

### 4.2 本研究で扱う問題の概要

本研究では，まず強化学習問題を有限の $N$ 期間（$N$ は既知の定数）の学習期間と有限の $T$ 期間（$T$ は既知の定数）の制御期間という 2 つの期間に分割して考える．そのもとで，統計的決定理論[2),5),13)]に基づいて制御期間の期待割引総収益の最大化を図る．最大化を行うにあたって，最適性の基準にはミニマックス基準，ベイズ基準[2),5),13)]などがあるが，本研究では Martin のアルゴリズム同様にベイズ基準を導入する．

状態集合 $S$，行動集合 $A$，学習期間の初期状態 $x_0$，制御期間の初期状態 $x_{N'}$，遷移確率行列 $p(s_j|s_i,a_k,\theta)$ のクラス，事前確率密度関数 $p(\theta)$，そして利得関数 $r(s_i,a_k,s_j)$ は既知と仮定する．また，真のパラメータ $\theta^*$ は未知，毎期の状態 $x_t$ は完全に観測されるものと仮定する．学習期間には毎期ごとに行動選択を行い，状態の遷移のみを観測し，収益は無視する．学習期間の最後（$N$ 期）には状態は $x_N$（遷移確率によって定まる確率変数）に遷移している．しかし，制御期間は $x_N$ ではなく，既知で固定の $x_{N'}$ から始まる．制御期間には毎期ごとに行動選択を行い，状態の遷移を観測するとともに，収益を得る．

学習期間の目的は，制御期間の期待割引総収益を最大化するような状態遷移のサンプルを得ることである．また，制御期間の目的は，実際に期待割引総収益を最大化することである．こうして制御期間の期待割引総収益を最大化するためには，そのような行動選択を行う学習期間と制御期間の政策 $\pi(\cdot,\cdots)$ を求めればよい．

### 4.3 有限の学習期間と有限の制御期間のもとでの定式化

以下で，ベイズ決定理論[2),5),13)]に基づく定式化を行う．

#### 4.3.1 効用関数

効用関数 $u(\theta,\pi(\cdot,\cdots),x^{N+T+2})$ は真のパラメータが $\theta$ であった場合に，決定関数として政策 $\pi(\cdot,\cdots)$ を用いて $x_0z_0\cdots x_{N+T}$ と遷移した場合に得られる割引総収益を示し，次式のように定義される．引数に行動

$z_t$ が含まれていないのは，毎期の行動は政策 $\pi(\cdot,\cdots)$ によって定まるためである．たとえば，$z_0=\pi(0,x_0)$ である．また，引数にパラメータ $\theta$ が含まれているが，$u(\theta,\pi(\cdot,\cdots),x^{N+T+2})$ の値は $\theta$ にはよらないものである．

$$u(\theta,\pi(\cdot,\cdots),x^{N+T+2}) = r(x_{N'},\pi(N,x_0z_0\cdots x_Nx_{N'}),X_{N+1}) + \sum_{i=N+1}^{N+T-1}\beta^{i-N}r(x_i,\pi(i,x_0z_0\cdots x_i),x_{i+1}). \quad (7)$$

上式の右辺が2つの項に分かれているのは，制御期間の初期状態 $x_{N'}$ が既知であるために，制御期間が学習期間の最後の遷移先 $x_N$ ではなく，$x_{N'}$ から始まっているためである．

#### 4.3.2 ベイズ期待効用関数

ベイズ期待効用関数 $E(u(\theta,\pi(\cdot,\cdots),x^{N+T+2})|\pi(\cdot,\cdots),p(\theta))$ は事前確率密度関数 $p(\theta)$ のもとで政策 $\pi(\cdot,\cdots)$ が用いられた場合の効用関数の期待値を示し，期待割引総収益に対応する．そのベイズ期待効用関数は式 (7) の効用関数をすべての遷移系列とパラメータ空間 $\Theta$ 全体で期待値をとることによって次式のように定義される．本研究では，この期待割引総収益を評価尺度とする．

$$E(u(\theta,\pi(\cdot,\cdots),x^{N+T+2})|\pi(\cdot,\cdots),p(\theta)) = \int_{\Theta}\sum_{x_1\cdots x_{N+T}} p(\theta)p(x_1\cdots x_{N+T}|\pi(\cdot,\cdots),x_0,x_{N'},\theta) u(\theta,\pi(\cdot,\cdots),x^{N+T+2})d\theta. \quad (8)$$

なお，右辺の条件付き確率は，条件部には学習期間と制御期間の初期状態も含まれており，これら2つの初期状態と政策とパラメータによる条件付きのもとでの遷移系列の発生確率を示している．この条件付き確率は次式のように遷移確率によって書き下される．

$$\begin{aligned} &p(x_1\cdots x_{N+T}|\pi(\cdot,\cdots),x_0,x_{N'},\theta)\\ &= p(x_1|x_0,\pi(0,x_0),\theta)p(x_2|x_1,\pi(1,x_0z_0x_1),\theta)\\ &\quad\cdots p(x_{N+1}|x_{N'},\pi(N,x_0z_0x_1\cdots x_Nx_{N'}),\theta)\\ &\quad\cdots p(x_{N+T}|x_{N+T-1},\\ &\qquad \pi(N+T-1,x_0z_0x_1\cdots x_{N+T-1}),\theta). \end{aligned} \quad (9)$$

#### 4.3.3 ベイズ決定

$BD(p(\theta))$ は事前確率密度関数 $p(\theta)$ のもとでのベイズ決定を示す．ベイズ決定は式 (8) のベイズ期待効用関数を最大にするような政策 $\pi(\cdot,\cdots)$ として次式のように定義される．

$$BD(p(\theta)) = \arg\max_{\pi(\cdot,\cdots)} E(u(\theta,\pi(\cdot,\cdots),x^{N+T+2})|\pi(\cdot,\cdots),p(\theta)). \quad (10)$$

式 (10) によるベイズ期待効用関数をベイズ基準のもとで最大，つまり，期待割引総収益をベイズ基準のもとで最大にするような政策を求めることが本研究の目的である．

なお，以下では最適と最大という言葉を式 (10) を満足するものとして，同一の意味として用いることにする．

また，式 (10) において，学習期間の長さを0としたのが Martin のアルゴリズムにおけるベイズ決定に相当する．すなわち，本研究は Martin のアルゴリズムの一般化となっている．

## 5. 提案アルゴリズム

### 5.1 基本最適アルゴリズム

#### 5.1.1 基本最適アルゴリズムの概要と最適性

本研究における目的を達成する最適解は式 (10) によって定義された．ここでは，DP（動的計画法，Dynamic Programming）[3] を用いることによって，実際にその最適解が求められる基本最適アルゴリズムを提案する．まずは，式 (10) による最適解が DP によって求められることを示す．

式 (10) を式 (3) と式 (9) を用いて書き下すと，次式のようになる．

$$\begin{aligned} &BD(p(\theta)) =\\ &\arg\max_{\pi(0,x_0)}\sum_{x_1}\bar{p}(x_1|x_0,\pi(0,x_0),x_0)\\ &\max_{\pi(1,x_0z_0x_1)}\sum_{x_2}\bar{p}(x_2|x_1,\pi(1,x_0z_0x_1),x_0z_0x_1)\cdots\\ &\max_{\pi(N,x_0z_0\cdots x_Nx_{N'})}\sum_{x_{N+1}}\bar{p}(x_{N+1}|x_{N'},\pi(N,x_0z_0\cdots x_Nx_{N'}),\\ &x_0z_0\cdots x_Nx_{N'})(r(x_{N'},\pi(N,x_0z_0\cdots x_Nx_{N'}),x_{N+1})\\ &+\beta\max_{\pi(N+1,x_0z_0\cdots x_{N+1})}\sum_{x_{N+2}}\bar{p}(x_{N+2}|x_{N+1},\pi(N+1,x_0z_0\\ &\cdots x_{N+1}),x_0z_0\cdots x_{N+1})(r(x_{N+1},\pi(N,x_0z_0\cdots\\ &x_{N+1}),x_{N+2})+\cdots\\ &+\beta\max_{\pi(N+T-2,x_0z_0\cdots x_{N+T-2})}\sum_{x_{N+T-1}}\bar{p}(x_{N+T-1}|x_{N+T-2},\\ &\pi(N+T-2,x_0z_0\cdots x_{N+T-2}),x_0z_0\cdots x_{N+T-2})\\ &(r(x_{N+T-2},\pi(N+T-2,x_0z_0\cdots x_{N+T-2}),x_{N+T-1})\\ &+\beta\max_{\pi(N+T-1,x_0z_0\cdots x_{N+T-1})}\sum_{x_{N+T}}\bar{p}(x_{N+T}|x_{N+T-1},\\ &\pi(N+T-1,x_0z_0\cdots x_{N+T-1}),x_0z_0\cdots x_{N+T-1})\\ &(r(x_{N+T-1},\pi(N+T-1,x_0z_0\cdots x_{N+T-1}),x_{N+T})\cdots)). \end{aligned} \quad (11)$$

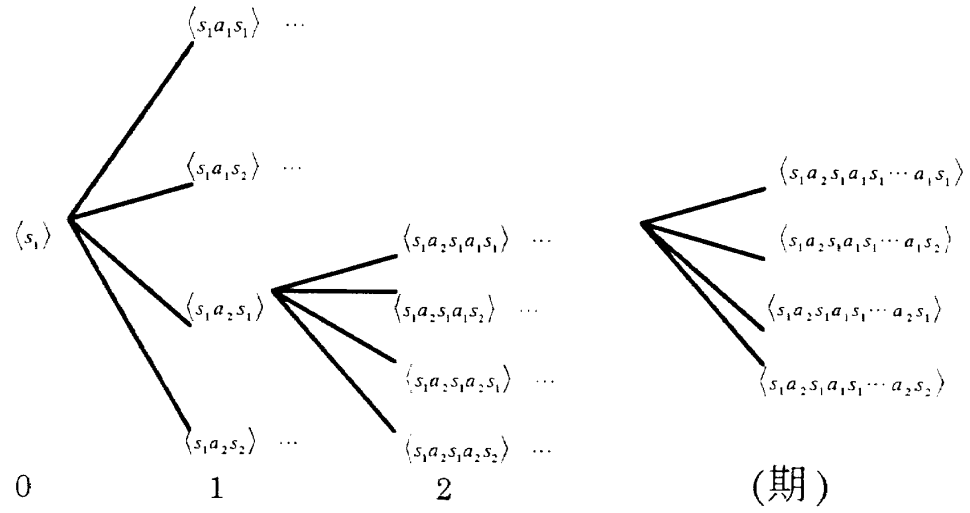


図 1 DP 木
Fig. 1 DP tree.

これは，式 (7) による効用関数を各期ごとに分解して，最大化を入れ子構造で表現している．$N$ 期に関しては，制御期間の初期状態 $x_{N'}$ を既知としているために，状態は $x_{N'}$，遷移系列は $x_0 z_0 \cdots x_N x_{N'}$ となっている．$N$ 期までは，学習期間のため収益は得られないので，単に確率の掛け合わせになっている．$N$ 期以降は，その期での収益とその期以降の最大の期待割引総収益の和になっている．

このように表現することによって，DP を用いて，まず，$N+T-1$ 期における 1 期間の最適化問題を解き，次に $N+T-2$ 期における 2 期間の最適化問題を先の 1 期間の最適解を利用して解くことができる．これを繰り返して，再帰的に式 (10) による最適解が求められる．

以上より，式 (10) を式 (11) のように書き下すことによって，ベイズ基準のもとでの最適解が DP を用いて求められる．

### 5.1.2 DP 木の導入

基本最適アルゴリズムを提案する前に，DP 木を導入する．これは，各ノードが遷移系列を示す木である．**図 1** に例を示す．なお，$S = \{s_1, s_2\}$，$A = \{a_1, a_2\}$，$x_0 = s_1$ とする．

式 (11) による最適解は，図 1 の葉のノード（右端のノード，$N+T-1$ 期のノード）における最適解を求め，次に葉から 1 つ溯った（1 つ左，$N+T-2$ 期の）ノードの最適解を葉のノードの最適解を利用して求めるというようにして，ルートのノード（左端のノード，0 期のノード）まで溯っていくことによって求めることができる．

### 5.1.3 基本最適アルゴリズムの提案

DP を用いて式 (11) を解く基本最適アルゴリズムの各ノードにおける行動選択を Step 1 から Step 3 に分けて以下に示す．なお，すでに DP 木は構成されているものとする．まず，Step 1 では制御期間の $N+T-1$ 期から $N+1$ 期の各ノードでの行動選択が決定される．制御期間の初期である $N$ 期の行動選択は Step 2 で決定される．これは，$N$ 期の状態が確率変数の $x_N$ ではなく，既知の制御期間の初期状態 $x_{N'}$ であるため，他の期とは分けていることによる．最後に Step 3 で学習期間の行動選択が決定される．

この Step 1 から Step 3 にわたる基本最適アルゴリズムによって，学習期間が有限の場合の，有限の制御期間の期待割引総収益がベイズ基準のもとで最大化されるのと同じことが DP を用いて再帰的に行われているので，式 (10) および式 (11) による最適解が得られる．

なお，このアルゴリズムにおいては，Martin によって提案されたアルゴリズム [7] と同様に事前確率密度関数としてベータ分布を仮定する．

**Step 1**：$t$ $(N < t \leq N+T-1)$ 期の各ノード $\langle x_0 z_0 \cdots x_t \rangle$ における行動選択は次式によって決定される．

$$z_t(x_0 z_0 \cdots x_t) = \arg\max_a \sum_{x_{t+1}} \bar{p}(x_{t+1} | x_t, a, x_0 z_0 \cdots x_t) \left( r(x_t, a, x_{t+1}) + \beta u'(x_0 z_0 \cdots x_t a x_{t+1}) \right), \tag{12}$$

ただし，$z_t(x_0 z_0 \cdots x_t)$ はノード $\langle x_0 z_0 \cdots x_t \rangle$ において選択するべき行動を示し，

$$u'(x_0z_0\cdots x_t) = \max_a \sum_{x_{t+1}} \bar{p}(x_{t+1}|x_t,a,x_0z_0\cdots x_t)\,(r(x_t,a,x_{t+1})+\beta u'(x_0z_0\cdots x_t a x_{t+1})), \quad (13)$$

かつ

$$u'(x_0z_0\cdots x_{N+T}) = 0. \quad (14)$$

式(12)中の $u'$ はそれまでに求められている1つ短い期間の最適解（正確には最適解である最適な行動選択による収益の期待値）に対応し，制御期間なので，その最適解に割引率を掛けたものにその期の収益を加えたものを，遷移確率行列を支配するパラメータの事後確率密度関数による遷移確率の加重平均で期待値をとって，さらにその最大化が行われている．式(13)ではさらに1つ長い期間の最適解が求められるときに利用される $u'$ として式(12)による最適解が保持されている．式(14)では，$N+T$ 期で終わりのため，$N+T$ 期以降の収益はないことを示している．

式(12)によって，$t$ 期から $N+T$ 期にかけて得ることのできる期待割引総収益を最大化するための制御期間中の $t$ 期における行動が決定される．

**Step 2**：$N$ 期の各ノード $\langle x_0z_0\cdots x_N\rangle$ における行動選択は次式によって決定される．

$$z_N(x_0z_0\cdots x_N) = \arg\max_a \sum_{x_{N+1}} \bar{p}(x_{N+1}|x_{N'},a,x_0z_0\cdots x_N x_{N'})\,(r(x_{N'},a,x_{N+1})+\beta u'(x_0z_0\cdots x_N x_{N'} a x_{N+1})), \quad (15)$$

ただし，

$$u'(x_0z_0\cdots x_{N+1}) = \max_a \sum_{x_{N+2}} \bar{p}(x_{N+2}|x_{N+1},a,x_0z_0\cdots x_{N+1})\,(r(x_{N+1},a,x_{N+2})+\beta u'(x_0z_0\cdots x_{N+1} a x_{N+2})). \quad (16)$$

式(15)によって，$N$ 期から $N+T$ 期にかけて得ることのできる期待割引総収益を最大化するための制御期間の初期状態における行動が決定される．

**Step 3**：$t$ $(0\le t\le N-1)$ 期の各ノード $\langle x_0z_0\cdots x_t\rangle$ における行動選択は次式によって決定される．

$$z_t(x_0z_0\cdots x_t) = \arg\max_a \sum_{x_{t+1}} \bar{p}(x_{t+1}|x_t,a,x_0z_0\cdots x_t)\,u'(x_0z_0\cdots x_t a x_{t+1}), \quad (17)$$

ただし，

$$u'(x_0z_0\cdots x_t) = \max_a \sum_{x_{t+1}} \bar{p}(x_{t+1}|x_t,a,x_0z_0\cdots x_t)\,u'(x_0z_0\cdots x_t a x_{t+1}), \quad (18)$$

かつ

$$u'(x_0z_0\cdots x_N) = \max_a \sum_{x_{N+1}} \bar{p}(x_{N+1}|x_{N'},a,x_0z_0\cdots x_N)\,(r(x_{N'},a,x_{N+1})+\beta u'(x_0z_0\cdots x_N x_{N'} a x_{N+1})). \quad (19)$$

学習期間は収益を得ないので，式(17)では $u'$（$N$ 期から $N+T$ 期にかけて得られる期待割引総収益）の最大化のみが行われている．つまり，式(17)では制御期間の期待割引総収益を最大化するための未知パラメータに関する情報収集のための行動が決定される．

#### 5.1.4 基本最適アルゴリズムの計算量

基本最適アルゴリズムの計算量はDP木の各ノードにおける行動選択の計算量の総和であり，各ノードにおける行動選択の計算量は定数オーダーである．よって，アルゴリズムの計算量はDP木における総ノード数に比例する．$node(t)$ を0期から $t$ 期のDP木における総ノード数とすると，0期から $t$ 期のDP木における総ノード数は次式で与えられる．

$$node(t) = \sum_{i=0}^{t} (|S||A|)^i. \quad (20)$$

すなわち，基本最適アルゴリズムの計算量は $t$ の指数オーダーである．メモリー量に関しては，DP木をそのまま記憶すると，DP木の総ノード数が $t$ の指数オーダーのため，指数オーダーである．DP木全体をそのまま記憶せずに，部分に分割して部分ごとに記憶しながらDPを解いていく方法もあるが，この場合，メモリー量は $O(t)$ になる．しかし，計算量はオーダー的には変化はないが，さらに大きくなる．

そこで，次節では計算量の軽減を目指して，この基本最適アルゴリズムの改良を行う．

### 5.2 改良最適アルゴリズム

前節において，式(10)および式(11)によって定義された最適解を実際に求めることができる基本最適アルゴリズムを提案した．しかし，基本最適アルゴリズムの計算量は指数オーダーである．そこで，本節においては基本最適アルゴリズムの改良を行い，新たに改良最適アルゴリズムを提案する．

#### 5.2.1 改良最適アルゴリズムの概要と最適性

改良最適アルゴリズムでは，基本最適アルゴリズムと同様に最適解が求められるという条件のもとで，そ

の計算量を軽減することが目的である．計算量はDP木のノード数に比例しているので，そのノード数をいかに軽減するかが重要である．

まずは，その基本的な考え方を示す．DPを用いた各ノードにおける行動選択は遷移確率行列を支配するパラメータの事後確率密度関数によって決まってくる．すなわち，基本的には，DP木の異なる（遷移系列が異なる）ノードであっても，同じ事後確率密度関数を有する場合には，それらのノードにおける最適な行動選択は同一のものとなる．よって，DP木において異なるノードとして表現されているが，実は同じ事後確率密度関数を有するノードを同一のノードとして表現することによって，総ノード数は軽減される．また，このようにしてノード数が軽減されても最適性は保持される．

### 5.2.2 遷移カウンターの導入

以下では，ノード数が軽減されるようなノードの表現形式を考えていくが，その準備として，遷移カウンターを導入する．今までは各ノードはそのノードに至るまでの遷移系列で区別されていたが，ここでは，その遷移系列中に含まれる個々の遷移（ある状態 $s_i$ においてある行動 $a_k$ が選択されてある状態 $s_j$ へ遷移するという1回の遷移）の回数で区別される．

学習期間において状態 $s_i$ において行動 $a_k$ を選択して状態 $s_j$ へ遷移した回数を $n_{s_i a_k s_j}$，制御期間において状態 $s_i$ において行動 $a_k$ を選択して状態 $s_j$ へ遷移した回数を $n'_{s_i a_k s_j}$，学習期間の遷移カウンターを $(n_{s_1 a_1 s_1}, n_{s_1 a_1 a_2}, \cdots n_{s_{|S|} a_{|A|} s_{|S|}})$，制御期間の遷移カウンターを $(n'_{s_1 a_1 s_1}, n'_{s_1 a_1 a_2}, \cdots n'_{s_{|S|} a_{|A|} s_{|S|}})'$ と表すと，遷移系列 $x_0 z_0 \cdots x_t$ から遷移カウンターを計算できる．

ただし，$t \leq N$ の場合には学習期間の遷移カウンターのみである．以下では，各ノードをこの遷移カウンターによって表す．遷移カウンターを学習期間と制御期間で分割して考えているのは，制御期間の初期状態を既知のものとしているためである．なお，$t$ 期の遷移カウンター $(n_{s_1 a_1 s_1}, n_{s_1 a_1 a_2}, \cdots n_{s_{|S|} a_{|A|} a_{|S|}})$ と $(n'_{s_1 a_1 s_1}, n'_{s_1 a_1 a_2}, \cdots n'_{s_{|S|} a_{|A|} s_{|S|}})'$ から状態 $x_t$ は次式によって求められる．

$t \leq N$ の場合，

$$x_t = \begin{cases} \arg\limits_{s_i \in S} n_{s_i \cdot \cdot} - 1 - n_{\cdot \cdot s_i} = -1, & s_i = x_0; \\ \arg\limits_{s_i \in S} n_{s_i \cdot \cdot} - n_{\cdot \cdot s_i} = -1, & s_i \neq x_0, \end{cases} \tag{21}$$

ただし，

$$n_{s_i \cdot \cdot} = \sum_{a_k, s_j} n_{s_i a_k s_j}, \tag{22}$$

かつ

$$n_{\cdot \cdot s_j} = \sum_{s_i, a_k} n_{s_i a_k s_j}. \tag{23}$$

$t > N$ の場合，

$$x_t = \begin{cases} \arg\limits_{s_i \in S} n'_{s_i \cdot \cdot} - 1 - n'_{\cdot \cdot s_i} = -1, & s_i = x_{N'}; \\ \arg\limits_{s_i \in S} n'_{s_i \cdot \cdot} - n'_{\cdot \cdot s_i} = -1, & s_i \neq x_{N'}, \end{cases} \tag{24}$$

ただし，

$$n'_{s_i \cdot \cdot} = \sum_{a_k, s_j} n'_{s_i a_k s_j}, \tag{25}$$

かつ

$$n'_{\cdot \cdot s_j} = \sum_{s_i, a_k} n'_{s_i a_k s_j}. \tag{26}$$

これは，式(22)および式(25)による状態 $s_i$ において行動選択が行われた回数と式(23)および式(26)による状態 $s_j$ へ遷移した回数を比較した場合，学習期間の初期状態 $x_0$ または制御期間の初期状態 $x_{N'}$ と状態 $x_t$ において差があることを利用している．

DPによる行動選択は，各ノードの有す事後確率密度関数に依存して決まるが，DP木において2つのノードの遷移カウンターが同一（制御期間においては，学習期間の遷移カウンターと制御期間の遷移カウンターがともに同一）であることは，その2つのノードが同一の事後確率密度関数を有し，かつ同じ状態にいることを示す．すなわち，その2つのノードにおける最適な行動選択は同一である．よって，2つのノードをマージしても最適性は保持される．このように同一の遷移カウンターを有すノードをマージすることによって，計算量の軽減が可能となる．なお，この遷移カウンターの考え方は，情報理論におけるタイプ[4]の考え方と同一である．

### 5.2.3 DPグラフの導入とそのノード数

次に，マージ操作によってノード数がどのくらい軽減されるかについて考えるが，その前にDPグラフを導入する．これは，前項におけるDP木のノード表現を遷移系列 $x_0 z_0 \cdots x_t$ から遷移カウンター $(n_{s_1 a_1 s_1}, n_{s_1 a_1 a_2}, \cdots n_{s_{|S|} a_{|A|} s_{|S|}})$ に変更し，さらにマージ操作を行うことによって木からグラフに表現形式を変更したものである．図**2**にDPグラフの例を示す．なお，$S = \{s_1, s_2\}$，$A = \{a_1, a_2\}$，$x_0 = s_1$ とし，簡単のため学習期間のノードのみの例であるためノードはすべて学習期間の遷移カウンターのみから成っている．

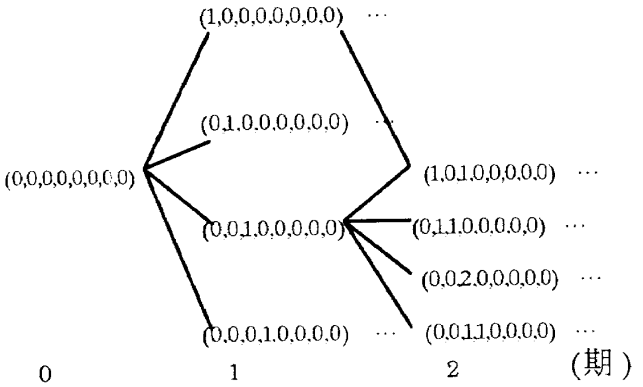


図 2 DP グラフ
Fig. 2 DP graph.

以下では，この DP グラフにおけるノード数について考える．学習期間中の $t$ 期の遷移カウンター $(n_{s_1a_1s_1}, n_{s_1a_1a_2}, \cdots n_{s_{|S|}a_{|A|}s_{|S|}})$ を構成する各 $n_{s_ia_ks_j}$ を見ると，0 から $t$ という $t+1$ 種類の数値をとりうる．よって，DP グラフにおける 0 期から $t$ 期の総ノード数 $node(t)$ は次式のような上界を有す．

$$node(t) < \sum_{i=0}^{t} (i+1)^{|S||A||S|}, \quad t \leq N. \tag{27}$$

同様に，制御期間中の $t$ 期について考えると，その上界は次式のように示される．上界が 2 つの項に分割されているのは，遷移カウンターが 2 つに分割されているためである．

$$node(t) < \sum_{i=0}^{N-1} (i+1)^{|S||A||S|} + (N+1)^{|S||A||S|} \sum_{i=0}^{t-N} (i+1)^{|S||A||S|}, \quad t > N. \tag{28}$$

これらの上界から分かるように，DP グラフにおけるノード数は多項式オーダーである．すなわち，ノードの表現形式を遷移系列から遷移カウンターに変更し，マージ操作を行うことによって，ノード数が指数オーダーから多項式オーダーに軽減された．

さらに，学習期間中の $t$ 期の場合，遷移カウンターの各 $n_{s_ia_ks_j}$ の合計は $t$ である．しかし，式 (27) の上界は各 $n_{s_ia_ks_j}$ の合計が $t$ でないものも含んでいるのでゆるい上界である．そこで，次式によってさらに厳しい上界を示す．

$$node(t) < \sum_{i=0}^{t} \begin{pmatrix} i+|S||A||S|-1 \\ |S||A||S|-1 \end{pmatrix}, t \leq N. \tag{29}$$

これは，$t$ 期の遷移カウンターの場合，一般的には $t$ 個の物を $|S||A||S|$ 個に分配する組合せの数で上界されることから導出したものである．同様に，制御期間中の $t$ 期の遷移カウンターの場合，次式によってさらに厳しい上界が示される．

$$node(t) < \sum_{i=0}^{N-1} \begin{pmatrix} i+|S||A||S|-1 \\ |S||A||S|-1 \end{pmatrix} + \begin{pmatrix} N+|S||A||S|-1 \\ |S||A||S|-1 \end{pmatrix} \sum_{i=0}^{t-N} \begin{pmatrix} i+|S||A||S|-1 \\ |S||A||S|-1 \end{pmatrix}, t > N. \tag{30}$$

また，学習期間と制御期間共通の下界が次式で示される．

$$\sum_{i=0}^{t} \begin{pmatrix} i+|A|-1 \\ |A|-1 \end{pmatrix} < node(t). \tag{31}$$

これは，$t$ 期の遷移カウンターの場合，$t$ 個の遷移を各行動（$|A|$ 個）に分配する組合せ数で下界としている．つまり，$t$ 個の遷移をどのように分配しようとも，それをさらに各行動のもとで $|S||S|$ 個に分配するので，遷移カウンターの数は必ず各行動への分配の組合せ数よりも大きくなる．

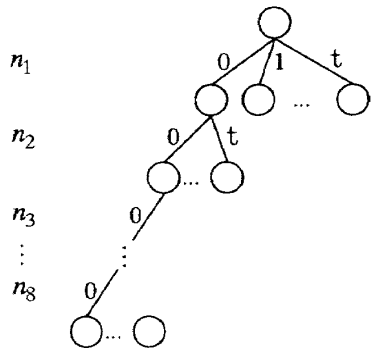


図 3 遷移カウンター木
Fig. 3 Transition counter tree.

DP グラフを作るときには，遷移カウンター（ノード）が生成されるたびに，すでに記録されている遷移カウンターの集合を探索して，その中に同一の遷移カウンターが存在しなければ，新たにその遷移カウンターを記録するという操作を行う．すなわち，すでに記録されている遷移カウンターを重複させて記録することはない．これが，マージ操作に対応している．

以下で，そのマージ操作に対応する遷移カウンターの探索手法を提案する．まずは，遷移カウンターの表現形式として，遷移カウンター木を導入する．これは，すでに生成されている DP グラフのノード（遷移カウンター）を木表現で保持しようというものである．図 3

に遷移カウンター木の例を示す．なお，$S=\{s_1,s_2\}$，$A=\{a_1,a_2\}$，$x_0=s_1$ とし，簡単のため学習期間の遷移カウンターのみの例である．

これは，各枝に遷移カウンターを構成する各 $n_{s_ia_ks_j}$ を次式によって書式を $n_{|S||A|\times(i-1)+|S|\times(k-1)+j}$ に変更したものが付与されている（深さ $i-1$ のノードから深さ $i$ のノードへの枝には $n_i$ 付与されている）．

$$n_{|S||A|\times(i-1)+|S|\times(k-1)+j}=n_{s_ia_ks_j}. \tag{32}$$

各葉が遷移カウンター（DP グラフのノード）に対応し，葉からルートに遡ることによって，遷移カウンターの値が分かる．

まず，学習期間の初期状態においては，すべての枝の値が 0 の葉のみが記録されている．その遷移カウンターから始めて，次期の遷移カウンターが 1 つずつ生成される．遷移カウンターが生成されるたびに，ルートから枝の値を用いて同一の遷移カウンターがすでに記録されていないか 2 分探索していく．同一のものが見つかれば，その遷移カウンターを新たに記録せず（マージ操作に対応），見つからなければ記録する．このとき，枝の値は左から昇順に記録する．制御期間については，学習期間の遷移カウンター木の葉のうち，枝の値の総和が $N$ である葉（学習期間の最後の遷移カウンターに対応）を新たにルートと見立てて，その下に制御期間の遷移カウンター木を作る．

### 5.2.4 改良最適アルゴリズムの提案

各ノードにおける行動選択に関する基本最適アルゴリズムと改良最適アルゴリズムの違いは，各ノードが遷移系列で表現されているか，遷移カウンターで表現されているかだけである．DP を用いて式 (11) を解く改良最適アルゴリズムの各ノードにおける行動選択を基本最適アルゴリズムと同様に，Step 1 から Step 3 に分けて，以下に示す．なお，すでに DP グラフは構成されているものとする．

また，ここでも Martin によって提案されたアルゴリズム [7] と同様に事前確率密度関数としてベータ分布を仮定する．

準備として，遷移カウンター $(n_{s_1a_1s_1},n_{s_1a_1s_2},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}})$（例として学習期間の場合）のもとでの状態 $s_i$ において行動 $a_k$ が選択されて状態 $s_j$ へ遷移する確率の，遷移確率行列を支配するパラメータの事後確率密度関数による加重平均を次式のように定義する．

$$\bar{p}(s_j|s_i,a_k,n_{s_1a_1s_1},n_{s_1a_1s_2},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}})=\int_\Theta p(\theta|n_{s_1a_1s_1},n_{s_1a_1s_2},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}})\,p(s_j|s_i,a_k,\theta)d\theta. \tag{33}$$

ただし，$p(\theta|n_{s_1a_1s_1},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}})$ は遷移カウンターが $(n_{s_1a_1s_1},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}})$ のときの事後確率密度関数を示す．

**Step 1**：$t$ $(N<t\le N+T-1)$ 期の各ノード $(n_{s_1a_1s_1},n_{s_1a_1s_2},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}},n'_{s_1a_1s_1},\cdots,n'_{s_{|S|}a_{|A|}s_{|S|}})$（学習期間の遷移カウンターと制御期間の遷移カウンターを合わせた表記を用いている）における行動選択は次式によって決定される．

$$\begin{aligned}&z(n_{s_1a_1s_1},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}},\cdots,n'_{s_{|S|}a_{|A|}s_{|S|}})\\&=\arg\max_a\sum_{x_{t+1}}\bar{p}(x_{t+1}|x_t,a,n_{s_1a_1s_1},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}},\cdots,n'_{s_{|S|}a_{|A|}s_{|S|}})(r(x_t,a,x_{t+1})\\&\quad+\beta u'(n_{s_1a_1s_1},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}},\cdots,n'_{x_tax_{t+1}}+1,\cdots,n'_{s_{|S|}a_{|A|}s_{|S|}})),\end{aligned} \tag{34}$$

ただし，

$z(n_{s_1a_1s_1},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}},\cdots,n'_{s_{|S|}a_{|A|}s_{|S|}})$ はノード $(n_{s_1a_1s_1},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}},\cdots,n'_{s_{|S|}a_{|A|}s_{|S|}})$ において選択するべき行動を示し，

$$\begin{aligned}&u'(n_{s_1a_1s_1},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}},\cdots,n'_{s_{|S|}a_{|A|}s_{|S|}})\\&=\max_a\sum_{x_{t+1}}\bar{p}(x_{t+1}|x_t,a,n_{s_1a_1s_1},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}},\cdots,n'_{s_{|S|}a_{|A|}s_{|S|}})(r(x_t,a,x_{t+1})\\&\quad+\beta u'(n_{s_1a_1s_1},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}},\cdots,n'_{x_tax_{t+1}}+1,\cdots,n'_{s_{|S|}a_{|A|}s_{|S|}})),\end{aligned} \tag{35}$$

かつ

$$\sum_{s_i,a_k,s_j}n'_{s_ia_ks_j}=T \tag{36}$$

の場合には，

$$u'(n_{s_1a_1s_1},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}},\cdots,n'_{s_{|S|}a_{|A|}s_{|S|}})=0. \tag{37}$$

**Step 2**：$N$ 期の各ノード $(n_{s_1a_1s_1},n_{s_1a_1s_2},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}})$（学習期間の遷移カウンターのみによる）における行動選択は次式によって決定される．

$$\begin{aligned}&z(n_{s_1a_1s_1},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}})\\&=\arg\max_a\sum_{x_{N+1}}\bar{p}(x_{N+1}|x_{N'},a,n_{s_1a_1s_1},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}})(r(x_{N'},a,x_{N+1})\\&\quad+\beta u'(n_{s_1a_1s_1},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}},\cdots,n'_{x_Nax_{N+1}}+1,\cdots,n'_{s_{|S|}a_{|A|}s_{|S|}})),\end{aligned} \tag{38}$$

ただし，

$$u'(n_{s_1a_1s_1},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}},\cdots,n'_{s_{|S|}a_{|A|}s_{|S|}})$$
$$=\max_a\sum_{x_{N+2}}\bar{p}(x_{N+2}|x_{N+1},a,n_{s_1a_1s_1},\cdots,$$
$$n_{s_{|S|}a_{|A|}s_{|S|}},\cdots,n'_{|S|a_{|A|}s_{|S|}})(r(x_{N+1},a,x_{N+2})$$
$$+\beta u'(n_{s_1a_1s_1},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}},\cdots,$$
$$n'_{x_{N+1}ax_{N+2}}+1,\cdots,n'_{s_{|S|}a_{|A|}s_{|S|}})). \quad (39)$$

**Step 3**：$t$ $(0 \leq t \leq N-1)$ 期の各ノード $(n_{s_1a_1s_1}, n_{s_1a_1s_2},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}})$ における行動選択は次式によって決定される．

$$z(n_{s_1a_1s_1},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}})$$
$$=\arg\max_a\sum_{x_{t+1}}\bar{p}(x_{t+1}|x_t,a,n_{s_1a_1s_1},\cdots,$$
$$n_{s_{|S|}a_{|A|}s_{|S|}})u'(n_{s_1a_1s_1},\cdots,n_{x_tax_{t+1}}+1,$$
$$\cdots,n_{s_{|S|}a_{|A|}s_{|S|}})), \quad (40)$$

ただし，

$$u'(n_{s_1a_1s_1},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}})$$
$$=\max_a\sum_{x_{t+1}}\bar{p}(x_{t+1}|x_t,a,n_{s_1a_1s_1},\cdots,$$
$$n_{s_{|S|}a_{|A|}s_{|S|}})u'(n_{s_1a_1s_1},\cdots,$$
$$n_{x_tax_{t+1}}+1,\cdots,n_{s_{|S|}a_{|A|}s_{|S|}})), \quad (41)$$

かつ

$$\sum_{s_i,a_k,s_j}n_{s_ia_ks_j}=N \quad (42)$$

の場合には，

$$u'(n_{s_1a_1s_1},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}})$$
$$=\max_a\sum_{x_{N+1}}\bar{p}(x_{N+1}|x_{N'},a,n_{s_1a_1s_1},\cdots,$$
$$n_{s_{|S|}a_{|A|}s_{|S|}})(r(x_{N'},a,x_{N+1})$$
$$+\beta u'(n_{s_1a_1s_1},\cdots,n_{s_{|S|}a_{|A|}s_{|S|}},\cdots,$$
$$n'_{x_{N'}ax_{N+1}}+1,\cdots,n'_{s_{|S|}a_{|A|}s_{|S|}})). \quad (43)$$

#### 5.2.5 改良最適アルゴリズムの計算量

改良最適アルゴリズムの計算量を考える．まず，DP グラフの各ノードにおける計算量は行動選択と定数回の遷移カウンター木の探索の計算量の和である．各ノードにおける行動選択の計算量は定数オーダーである．

遷移カウンター木の探索については，学習期間の場合，遷移カウンター木の各ノードでの 2 分探索は $t$ 期であれば，枝は最大で $t+1$ 本なので，計算量は $O(\log t)$ である．木の深さは定数 $|S||A||S|$ なので，1 つの遷移カウンターの探索の計算量は $O(\log t)$ である．制御期間の探索に関しても学習期間同様にその計算量は $O(\log t)$ である．つまり，各ノードにおける遷移カウンター木の探索の計算量は $O(\log t)$ である．

よって，DP グラフの各ノードにおける行動選択と遷移カウンター木の探索の計算量は，それぞれ定数オーダーと $O(\log t)$ である．改良最適アルゴリズム全体での計算量を考えてみると，これが $t$ の多項式オーダーのノード数分行われるので，全体では $t$ の多項式オーダーである．

以上のように，計算量は基本最適アルゴリズムにおいて $t$ の指数オーダーであったものが $t$ の多項式オーダーに軽減された．

メモリー量は，DP グラフ全体をそのまま記憶すると，DP グラフ，遷移カウンター木ともにそのノード数が $t$ の多項式オーダーのため，$t$ の多項式オーダーである．

## 6. まとめ

本研究では，強化学習問題を学習と制御に明確に分けて解くために，学習期間と制御期間からなる強化学習問題を扱った．従来から学習期間と制御期間に分割された強化学習問題は研究されている．それらの研究では，学習期間が無限の場合に真の最適政策への収束が保証されているが，学習期間が有限の場合には制御期間の収益最大化の厳密な保証はない．

そこで，本研究では有限で固定の学習期間と有限で固定の制御期間からなる強化学習問題を扱い，制御期間の期待割引総収益の最大化を行った．最適性の基準としては，有限で固定の制御期間のみからなる強化学習問題を扱った Martin のアルゴリズムと同様にベイズ決定理論を導入した．

まず，ベイズ決定理論に基づいて制御期間の期待割引総収益の最大化の定式化を行った．次に，実際にその最適解が求められる基本最適アルゴリズムの提案を行った．しかし，その基本最適アルゴリズムの計算量が $t$ の指数オーダーであったため，基本最適アルゴリズムの改良を行った．改良最適アルゴリズムは基本最適アルゴリズムと同様に最適解が求められるうえに，その計算量は $t$ の多項式オーダーにまで軽減されている．

また，基本最適アルゴリズムの学習期間の長さを 0 にしたものは Martin のアルゴリズムに相当する．よって，Martin のアルゴリズムもまたその計算量は $t$ の指数オーダーであり，改良最適アルゴリズムの学習期間の長さを 0 にすることによって，Martin のアルゴリズムの計算量を $t$ の多項式オーダーにまで軽減することも可能である．

**謝辞** 本研究を行うにあたり，ご討論，ご助言いただいた早稲田大学・新家稔央氏，後藤正幸氏，中澤真氏，NTT コミュニケーション科学研究所・向内隆文氏ならびに早稲田大学・平澤，松嶋両研究室の各氏に深く感謝申し上げます．本研究の一部は文部省科学研究費基盤 A. 展開（No.07558168）の補助による．

## 参 考 文 献

1) Barto, A.G., Bradtke, S.J. and Singh, S.P.: Learning to act using real-time dynamic programming, *J. of Artificial Intelligence*, Vol.72, No.1-2, pp.81–138 (1995).
2) Berger, J.: *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag (1985).
3) Bertsekas, D.P.: *Dynamic Programming and Stochastic Control*, Academic Press (1976).
4) Blahut, R.E.: *Principles and Practice of Information Theory*, Addison-Wesley (1987).
5) Ferguson, T.S.: *Mathematical Statistics*, Academic Press (1967).
6) 金子哲夫：マルコフ決定理論入門，槙書店 (1973).
7) Martin, J.J.: *Bayesian Decision Problems and Markov Chains*, John Wiley & Sons (1967).
8) Martin, L.P.: *Markov Decision Processes*, John Wiley & Sons (1994).
9) 宮崎和光，山村雅幸，小林重信：k-確実探査法 強化学習における環境同定のための行動選択戦略，人工知能学会誌， Vol.10, No.3, pp.454–463 (1995).
10) 宮崎和光，山村雅幸，小林重信：l-確実探査法 エージェントによる環境同定のための行動選択戦略，人工知能学会誌， Vol.11, No.5, pp.804–808 (1996).
11) 宮崎和光，山村雅幸，小林重信：MarcoPolo 報酬獲得と環境同定のトレードオフを考慮した強化学習システム，人工知能学会誌， Vol.12, No.1, pp.78–89 (1997).
12) 佐藤光男，阿部健一，竹田 宏：ベイズ決定手法を用いた未知パラメータを含むマルコフ決定過程の漸近的性質，電子通信学会論文誌，Vol.J61-D, No.1, pp.1–8 (1978).
13) 繁桝算男：ベイズ統計入門，東京大学出版会 (1985).
14) 高橋幸雄，森村英典：マルコフ解析，日科技連 (1979).
15) Watkins, C.J.C.H. and Dayan, P.: Q-Learning, *Machine Learning*, Vol.8, pp.279–292 (1992).

(平成 9 年 3 月 31 日受付)
(平成 10 年 1 月 16 日採録)

**前田 康成**

平成 7 年早稲田大学理工学部工業経営学科卒業．平成 9 年同大学院理工学研究科修士課程修了．同年，日本電信電話（株）入社．現在，NTT 情報通信研究所勤務．在学中，機械学習，特に強化学習の研究に従事．

**浮田 善文**

平成 6 年早稲田大学理工学部工業経営学科卒業．平成 8 年同大学院理工学研究科修士課程修了．同年，同大学院理工学研究科博士後期課程入学，現在に至る．機械学習の研究に従事．

**松嶋 敏泰**（正会員）

昭和 53 年早稲田大学理工学部工業経営学科卒業．昭和 55 年同大学院理工学研究科博士前期課程修了．同年，日本電気（株）入社．昭和 61 年早稲田大学大学院理工学研究科博士後期課程入学．平成元年横浜商科大学講師．平成 4 年同大学助教授．平成 5 年早稲田大学理工学部工業経営学科助教授．平成 9 年早稲田大学理工学部経営システム工学科教授，現在に至る．知識情報処理および情報理論とその応用に関する研究に従事．工学博士．IEEE，電子情報通信学会，人工知能学会，情報理論とその応用学会等各会員．

**平澤 茂一**（正会員）

昭和 36 年早稲田大学理工学部数学科卒業．昭和 38 年同電気通信学科卒業．同年，三菱電機（株）入社．昭和 56 年早稲田大学理工学部工業経営学科（現在経営システム工学科）教授，現在に至る．情報理論とその応用，データ伝送方式，ならびに計算機応用システムの開発などの研究に従事．工学博士．昭和 60 年ハンガリー科学アカデミー，昭和 61 年伊トリエステ大学客員研究員．平成 5 年電子情報通信学会小林記念特別賞，業績賞受賞．IEEE Fellow，電子情報通信学会，情報理論とその応用学会，人工知能学会，OR 学会，日本経営工学会等各会員．