適応制御過程

蔵野　正美

1. はじめに

個人あるいは団体の意思決定は，ほとんどの場合未来に向けての行動選択と考えられる．未来における結果は本質的に確定的でありえないから，不確実性のもとでの決定問題を考えねばならない．たとえば幼稚園児の遠足では，ご馳走弁当の下ごしらえ等を前の晩に用意する．しかし，もし当日雨が降って遠足が中止になるとご馳走がふいになってしまう．これは意思決定の損失と考えられよう．逆の場合もありえる．すなわち，明日は雨だと思って，弁当の用意をしなかったが，当日雨がやんで遠足が可能になった場合，園児は落胆し主婦の嘆きはさぞ大変なものだろう．賢い主婦は明日雨が降るか否かを予測してご馳走を作るかどうか，損失を考えてその内容および経費の掛け具合を決めるであろう．このように，われわれの行動決定には多くの場合不確実性のもとでなされるが，これに対処する1つの方法として，不確実性を確率（法則）で測り，各行動に対する効用の期待値を最大にするように行動選択を行なうことが考えられる[13]．しかし，この場合不確実性のありようの表現である確率法則をいかにして定めるかが大問題である．遠足の例では，テレビの天気予報を見るとか，友人の意見を聞くとかしてよりよい決定を行なうためにいろいろと明日の天気についての情報を集めるであろう．新しい情報がはいるたびに，さらに新しい情報を得るために行動をおこすべきか，もしおこすとしたらどのような行動をとるべきか，あるいは，手持ちの情報で満足して最終的な判断を下すべきかと主婦の気持は動揺して，いくつかの屈折を経たのち最終決定がなされる．この身近な例でもわかるように，未知の要素が含まれている最適化問題では未知の要素についての有益な情報をいかにして収集し，また得られた情報を決定過程の中に組み込み，いかにして最適化をはかってゆくかの問題がとり扱われる．いわゆる**適応型**の問題といわれるものである．このように考えてくると，ORでとり扱われる数学モデル，たとえば在庫モデル，取替モデル，ゲームモデルなどにはすべて適応型のモデルのあてはまる場合が非常に多い．しかもこれは現実をより反映したモデルといえよう．ところが適応型の問題は情報をとり扱うため，一般的にモデルが複雑になり解析的にも数値的に最適解を求めるのが困難である．これに対処する一方法として，ベルマン(Bellman)の開発した**DP**［1］は最有力である．DPの手法があるからこそ適応型の問題が解けるようになったと言えば過言すぎるであろうか．ここでは，適応型の問題を適応制御過程あるいはマルコフ決定過程として定式化し，DPの方法を使って解く方法を解説する．さらに応用範囲の広いマルコフ過程の適応制御をとりあげて情報処理と最適化とに関するいくつかのアプローチの方法と結果を紹介する．

くらの　まさみ　千葉大学　教育学部

2. マルコフ決定過程とDP

たばこ製造会社は年度末の自社製品のシェアにもとづいて次の年度の広告費を決める．ある計画期間(会計年度を1期間として，たとえばN期間)にわたって総利益を最大にするためには，各年度にどのくらいの広告費を支出すればよいであろうか．タバコのシェアは年々変動するがその年度末のシェアを変化する状態と考え，支出される広告費を行動（action）としてこの問題は逐次決定過程，特に**マルコフ決定過程**（Markov Decision Process, **MDP**)によって定式化される．その前にMDPの定義を与えておく．n期でのシステムの状態i_nは状態空間$S=\{1,2,\cdots,L\}$の1点として表わされ，行動a_nは行動空間$A=\{1,2,\cdots,K\}$から決定者によって選択されるものとする．たとえば，たばこ製造会社の問題ではシェアが15%以上（未満）である状態を1(2)で表わせば，$S=\{1,2\}$となる．状態の推移(dynamics)に関しては定常性とマルコフ性を仮定する．すなわち，推移確率行列$q=(q_{ij}{}^k)$，ただし$q_{ij}{}^k\geqq 0$，$\sum_{j=1}^{L} q_{ij}{}^k=1$が与えられていて，Prob $(i_{n+1}=j|i_0,a_0,\cdots,i_n=i,a_n=k)=q_{ij}{}^k, i,j\in S, k\in A$が成り立つものとする．また与えられた利得行列$r=(r(i,k))$に対して，$i_n=i$で行動$a_n=k$を選ぶと利得$r(i,k)$が発生する．各期の行動$a_n$を選択する規則を政策という．一般に$a_n$は$n$期以前の履歴$h_n=(i_0,a_0,\cdots,i_n)$に依存して確率的に選ばれるが，特に$n$期の状態$i_n$のみに依存して確率1である行動を選ぶ政策を確定的なマルコフ型の政策といい，SからAへの写像f_nの組$\pi=(f_0,f_1,\cdots,f_N)$で表わされる．この政策を使うとn期では$i_n=i$のとき確率1で$i_n=f_n(i)$が選択されることになる．また$f=f_n(n\geqq 0)$のとき定常政策といい，簡単のためにfで表わす．政策πを使ったときの総期待利得は初期状態$i_0=i$の関数として次のように表わされる

(2.1) $V_N(i,\pi)=\sum_{n=0}^{N} E_\pi[\beta^n r(i_n,a_n)|i_0=i]$.

ただし，$0<\beta\leqq 1$は割引き率．

(2.1)を最大にする最適政策を求める問題を考えてみよう．目的関数(2.1)には加法性があるのでDPの**最適性の原理**[1]を適用して解くことができる．

$V_n(i)=n$期間問題で初期状態がiのとき，最適政策のもとでの最大総期待利得

とすると次の再帰関係式を満足する．

(2.2) $$V_0(i)=\max_{k\in A} r(i,k)$$
$$V_n(i)=\max_{k\in A}\{r(i,k)+\beta\sum_{j=1}^{L} q_{ij}{}^k V_{n-1}(j)\}$$
$$i\in S, n=1,2,\cdots,N$$

ここで，各$i\in S$に対して(2.2)の右辺の最大値を与える1つの行動を$f_n(i)$として，$f_n: S\to A$を定義すると$\pi^*=(f_N,f_{N-1},\cdots,f_0)$は最適政策となる．

無限期間（$N=\infty$）での基準として

(i)割引き率β $(0<\beta<1)$のある割合（割引き最適基準）

(2.3) $$V_\beta(i,\pi)=\sum_{n=0}^{\infty} E_\pi[\beta^n r(i_n,a_n)|i_0=i]$$

(ii)長期間における1期当りの平均期待利得（平均最適基準）

(2.4) $$g(i,\pi)=\lim_{N\to\infty}\inf(N+1)^{-1}\sum_{N=0}^{N} E_\pi [r(i_n,a_n)|i_0=i]$$

たとえば，表1で示すデータのもとでの平均最適基準における最適政策は状態1では2 (low advertising)，状態2では1 (high advertising)をとる定常政策となり，そのときの1期当りの平均利得$=56.37$となる．

MDPはその構造の一般性により，ORでとり扱う決定過程，統計学の逐次解析，最適制御など応用範囲がきわめて広く，この20年間で膨大な量の研究論文が発表されている．MDPの一般理論，計算アルゴリズム，応用に関するsurveyはそれぞれ[4, 14, 9]に与えられている．またMDPの日本語の教科書として[5, 6, 18]をあげておく．

3. 適応型の決定モデル

今，2台のスロットマシンI，IIがある．マシンIを使えば確率rで1ドルが得られ，マシンII

表 1　たばこ製造会社のデータ

状態 i	行動 k	推移確率 q_{i1}^k	推移確率 q_{i2}^k	利得 r_i^k
1. シェアが15%以上	1	0.7	0.3	68
	2	0.5	0.5	80
2. シェアが15%未満	1	0.6	0.4	25
	2	0.33	0.67	33.2

ただし，行動 1 (2)：高い(安い)広告費の支出

を使えば確率 p で 1 ドルが得られるとする．各期ではマシンⅠ，Ⅱのどちらかを選んで使う．このとき，N 期間に得られる総期待金額を最大にするためには，各期でマシンⅠ，Ⅱのどちらを選んで使えば良いであろうか．もし r と p の値が既知であれば．1 ドルが出てくる確率 r，p の大きいほうのマシンを毎期使えばよい．しかし，r，p の値が未知の場合はどうふるまえばよいであろうか．これは逐次実験計画の中で特に**2 腕の盗賊**の問題 (two-armed bandit problem) [2, 17, 20] と呼ばれている．今，簡単のために r の値は既知で p の値が未知であるとしよう．p の値が未知といっても，第 1 期目の選択を行なう時に決定者は p に対するなんらかの情報・予備知識（これを**初期情報**という）をもっていることが考えられる．たとえば，マシンⅡを使ったことのある友人から情報の提供を受けたり，宣伝文を読んだりして初期情報を豊富にするように努めるかもしれない．またなんの予備知識をもち合せてない場合でも広く解釈して，それ自体がまた 1 つの初期情報と考えられる．今，マシンⅡを最初から n 期のあいだ，つづけて使ったとする．このとき，n 回の実験結果1100…10から得られる情報が初期情報に追加され，これらの情報をもとにして $n+1$ 期の選択がなされる．ただし 1 (0) は成功(失敗)を表わす．

一般に未知パラメータを含む決定過程をＤＰで定式化して解くためには，決定に役立つあらゆる場合の情報がある集合 $\mathcal{S}_1$ の一点として表わされかつ過程の進行にともない追加される情報が $\mathcal{S}_1$ 上の変換あるいは推移として表わされる必要がある．情報をこのような形で表わしたものを**情報様式** (Information Pattern) という．この場合 $\mathcal{S}_1$ は過程の状態空間の一部を構成し，追加情報による情報様式の改訂の方法が状態の推移(確率)法則を規定することになる．

以上を考慮して典型的な**適応制御過程**を定式化してみる．過程の状態はその期までの系の状態 p とその期までの未知パラメータに関する情報の表現である情報様式 P の対 (p, P) で表わされ，(p, P) のとりうる値の全体を $\mathcal{S}$ とする．状態 $(pP) \in \mathcal{S}$ に対して行動空間 $\mathcal{A}$ から 1 つの行動 $a \in \mathcal{A}$ を選択すると利得 $R((p, P), a)$ が発生する．またこのとき系の状態の推移と追加情報が得られるがこのメカニズムが状態空間 $\mathcal{S}$ の上の推移確率法則 $Q(\cdot|(p\,P), a)$ で表わされると仮定する．

明らかに上で定義された決定過程は $\mathcal{S}$ を状態空間にもつＭＤＰであり，ＤＰの再帰関係式(2.2)を解くことによって最適な適応政策が求められる．現実の具体的問題に対しては，情報様式の選択およびその改訂の方法が問題になるが，これには統計学の推測論，特に十分統計量が利用される．通常，未知パラメータの“もっともらしさ”の程度を表わす確率分布で情報を記述して，追加情報をベイズの定理によって事前分布から事後分布に改定する方法がよく用いられる．これを**ベイズ** (Bayes) モデルと呼ぶ．スロットマシンの問題を Bayes モデルで解いたのが図 1 である．ただし $N=6$，$r=0.6$，初期情報は一様分布．このとき最大期待利得は3.72ドルとなり，常にⅠを使うときの期待利得 3.6ドルより 0.12ドルと高くなっている．2 腕の盗賊の問題については，医学的な治療法の選択の問題などに応用されている([15])．

4. マルコフ過程の適応制御

適応型の逐次決定モデルでは，情報様式の選択およびその改訂の方法が重要であることはすでに述べたが，そのアプローチの方法の違いによって，いろいろなタイプの適応制御モデルが考えられ

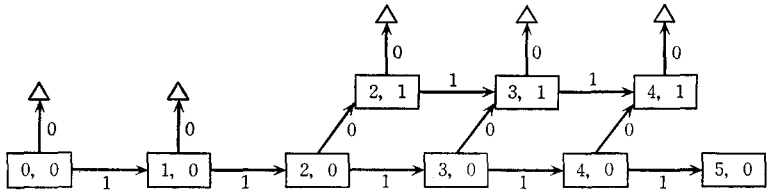

図 1 最適政策$N=6$, $r=0.6$
ただし$\boxed{i,\ j}$はIIを使用，$i(j)$は成功(失敗)の回数，△では以後Iのみ使用

る．ここでは，MDPで特に状態の推移確率法則に未知のパラメータが含まれる場合，すなわち，m次元ユークリッド空間の部分集合Θをパラメータ空間として，各$\theta\in\Theta$に対して1つのMDPの推移確率法則$q(\theta)=(q_{ij}{}^k(\theta))$が対応しており，$\theta$の値が未知の場合についての適応制御について考えてみよう．この場合，(2.3)あるいは(2.4)で定義されたV_β, gは未知パラメータの値θに依存するので，θの関数として，$V_\beta(i,\pi,\theta)$, $g(i,\pi,\theta)$と表わされる．

4.1 ベイズモデル

すでに述べたように，初期情報をパラメータ空間Θの上の事前確率分布で表わし，情報の蓄積を事後分布で記述する方法がベイズモデルである．相当する適応制御過程のn期の状態はi_nとn期の履歴のもとでのΘの上の事後分布ξ_nの対(i_n,ξ_n)で表わされ，したがって状態空間は$S\times P(\Theta)$となる．ただし$P(\Theta)$はΘの上の確率分布の全体．事前分布$\xi\in P(\Theta)$に対して，状態iで行動kを選択し，次の期に状態がjに移行したという情報のもとでの事後分布を$T_{ij}{}^k\xi$と表わすと，DPの再帰関係式(2.2)は次のようになる．$n\geqq 1$, $(i,\xi)\in S\times P(\Theta)$に対して

$$(4.1)\quad V_n(i,\xi)=\max_{k\in A}\cup_k V_{n-1}(i,\xi),$$

ただし $V_0(i,\xi)=\max_{k\in A} r(i,k)$

$$\cup_k u(i,\xi)=r(i,k)+\beta\sum_{j\in S}\int q_{ij}{}^k(\theta)\,u(j,T_{ij}{}^k\xi)\xi(d\theta)$$

(4.1)で$V_\beta(i,\xi)=\lim_{n\to\infty} V_n(i,\xi)$ (ただし$0<\beta<1$)とおけば，割引最適基準の最適方程式

$$(4.2)\quad V_\beta(i,\xi)=\max_{k\in A}\cup_k V_\beta(i,\xi)$$ を得る．

今，$f(i,\xi)=\arg\max_{k\in A}\cup_k V_\beta(i,\xi)$とする時，$n$期の行動として$a_n=f(i_n,\xi_n)$を選ぶ定常政策が**ベイズ最適**となることが知られている．すなわち

$$V_\beta(i,\xi)=\max_\pi\int V_\beta(i,\pi,\theta)\xi(d\theta)$$
$$=\int V_\beta(i,f,\theta)\xi(d\theta).$$

しかし，ベイズ最適政策を求めるためには，結局(4.2)式を解く必要があり困難が予想される．そこで，簡単でとり扱いやすいBayesian equivalent ruleなどの近似的な政策が提案されその性質が調べられている[19]．

次に未知パラメータに関する情報の収集という立場からベイズ最適政策を検討してみよう．

例 ([8]) $S=A=\{1,2\}$, $r_1{}^1=r_1{}^2=1$, $r_2{}^1=r_2{}^2=0$, 推移行列qは図2のように，$q_{22}{}^2=\theta=1-q_{21}{}^2$のみが未知で他は既知とする．この場合$\Theta=[0,1]$で，初期分布を適当にとるとベイズ最適政策のもとでは，確率1で$a_n=1$ $(n\geqq 0)$となることが最適方程式(4.2)から示される．これは状態2において行動1を常にとることを意味しており，未知パラメータθに対してなんらの情報も得られないことになる．そのために，たとえばθが0に十分近いとき明らかに状態2では行動2をとるのが最適であるのにベイズ政策のもとでは2を選ばないで1を選ぶという結果になっている．

このようなベイズ政策の欠点は漸近的最適性の概念，あるいはForced Choice Circleなどの手法が導入されて解決されている([8, 16, 19])．

4.2 ノンベイズモデル

事前分布の存在を仮定しない，いわゆるNon-Bayesian的な方法を考察してみよう．各$\theta\in\Theta$および任意の政策πと初期状態$i\in S$に対して

$$(4.3)\quad g(i,\pi^*,\theta)\geqq g(i,\pi,\theta)$$

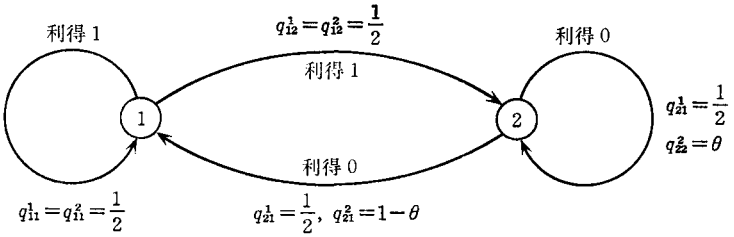

図 2 未知パラメータをもつ推移確率行列

が成り立つ π^* を平均最適な適応政策と呼び，π^* の構成法の問題を検討してみよう．1つの考え方として，蓄積された情報をもとにして未知のパラメータを推定し，この推定値がパラメータの真値とみなしたとき，最適となる行動を選択する方法，すなわち**推定と制御の原理**（[7, 12]）を挙げることができる．今，この考え方によって1つの適応政策を構成してみよう．推定量として最尤推定量（MLE）を利用することにする．

(4.4) $\hat{\theta}_n \equiv \arg\max_{\theta\in\Theta} L_n(\theta, h_n)\ (n\geqq 0)$

ここで $L_n(\theta, h_n)=\sum_{t=0}^{n-1}\log q(i_{t+1}|i_t, a_t)$

ただし $q(j|i,k)=q_{ij}^k$, $h_n=(i_0, a_0, \cdots, i_n)$

今，各 $\theta\in\Theta$ に対して $g(i, f[\theta], \theta)=\sup_{\pi} g(i, \pi, \theta)$ が成り立つ最適定常政策 $f[\theta]: S\to A$ が存在すると仮定しよう．このとき n 期において履歴 h_n から (4.4) によって $\hat{\theta}_n$ を求め，$\hat{\theta}_n$ をパラメータの値とみなしたときに最適となる行動 $a_n=f[\hat{\theta}_n](i)$ を選択する政策 $\pi^*=(f[\hat{\theta}_1], [\hat{\theta}_2], \cdots)$ が考えられる．この適応政策 π^* を使った時ある条件のもとでMLEの一致性，すなわち確率1で $\hat{\theta}_n$ がパラメータの値 θ_0 に収束すること，および π^* が平均最適となることが示されている（[7, 12]）．

4.3 学習モデル

各期に得られる情報をいわゆる reward-penalty 型の学習方式（[11]）と value-iteration（[3]）で処理する方法について述べよう．n 期での行動 a_n に対する条件付き確率分布を

$$\tilde{\pi}_n(k|i)=\mathrm{Prob}(a_n=k|i_0, a_0, \cdots, i_n=i)\ (k\in A)$$

で表わす．このとき $\tilde{\pi}_n$ が n を増大させるといくらでも平均最適な政策に近づく $\{\tilde{\pi}_n\}$ に対する学習アルゴリズムを求める．正の推移確率行列 q の全体 $Q=\{q=(q_{ij}^k)\,|\,q_{ij}^k>0, \sum_{j=1}^{L} q_{ij}^k=1\}$ をパラメータ空間として，q の推定量としてMLE $q(n)=(q_{ij}^k(n))$ を使う．ここで，$q_{ij}^k(n)=n_{ij}^k/\sum_{j\in S} n_{ij}^k$，ただし n_{ij}^k は n 期の履歴 $h_n=(i_0, a_0, \cdots, i_n)$ において $i_t=i$, $a_t=k$, $i_{t+1}=j$ となる t の個数．初期情報 $q(0)\in Q$ と $\lambda_n\in(0,1)$ なる数列 $\{\lambda_n\}$ を適当に選んで n 期の情報様式を $\tilde{q}(n)=\lambda_n q(0)+(1-\lambda_n)q(n)$ で表わす．各 $k\in A$ と $q\in Q$ に対して，L 次元ユークリッド空間 $\boldsymbol{R}^L$ の作用素 $U^k[q]=(U_1^k[q], \cdots, U_L^k[g])$ を

(4.5) $U_i^k[q]\,\boldsymbol{u}=r(i,k)+\sum_{j\in S}(q_{ij}^k-\eta(q))u_j,$

$\boldsymbol{u}=(u_1, \cdots, u_L)\in \boldsymbol{R}^L$

ただし $\eta(q)=\min_{i,\,j\in S,\,k\in K} q_{ij}^k$

で定める．このとき $\eta(q)>0$ であるから，作用素 $U[q]=\max_{k\in A} U^k[q]$ は縮小写像となる．今，不動点を $\boldsymbol{u}^*=(u_1^*, u_2^*, \cdots, u_L^*)$ とすれば，$\boldsymbol{u}^*=U[q]\boldsymbol{u}^*$ が成り立ち，この式において $\eta^*=\eta(q)\sum_{j\in S} u_j^*$ とおけば，平均最適基準のもとでの最適方程式を得る：

(4.6) $u_i^*=\max_{k\in A}\{r(i,k)-\eta^*+\sum_{j\in S} q_{ij}^k u_j^*\}\ (i\in S)$

各 $i\in S$ に対して，上式の右辺を最大にする $k\in A$ の全体を $A_i^*[q]$ で表わすと $f(i)\in A_i^*[q]$ $(i\in S)$ となる任意の定常政策 f は q に対して定まるMDPの平均最適となり，η^* はそのときの最適平均期待利得となる（[5, 6, 18]）．今，推定量 $\tilde{q}(n)$ を使って L 次元ベクトルの列 $\{\tilde{\boldsymbol{V}}(n)=(\tilde{V}_1(n), \tilde{V}_2(n), \cdots, \tilde{V}_L(n))\}$ を逐次決めてゆくことにする（value-iteration）．

(4.7) $\tilde{V}_i(0)=0$

$$\tilde{V}_i(n+1)=\max_{k\in A} U^k[\tilde{q}(n)]\tilde{\boldsymbol{V}}(n)\ \ (n\geqq 0)$$

次に $b_0=1$, $b_n>b_{n+1}>0 (n\geqq 0)$ なる数列 $\{b_n\}$ に対して，$\phi(b_n)=b_{n+1}(n\geqq 0)$ を満たす増加関数

ϕ : [0, 1]→[0, 1] によって，$\bar{\pi}_{n-1}$ を $\bar{\pi}_n$ に次のように改訂する（学習アルゴリズム）．(4.7) の右辺を最大にする任意の k を $\tilde{k}_{n+1}(i)$ で表わす．各 $i \in S$ に対して，$\tilde{k}_{n+1}(i)=k_i$ ならば

$$(4.8)\quad \bar{\pi}_n(k_i|i)=1-\sum_{l\neq k_i}\phi(\bar{\pi}_{n-1}(l|i))$$
$$\bar{\pi}_n(l|i)=\phi(\bar{\pi}_{n-1}(l|i))\quad (l\neq k_i).$$

このアルゴリズムでは，$\bar{\pi}_n(k_i|i)>\bar{\pi}_{n-1}(k_i|i)$，$\bar{\pi}_n(l|i)<\bar{\pi}_n(l|i)$ $(l\neq k_i)$ が成り立つので(4.8)は一種の reward-penalty 型の学習方式となっている．(4.7)と(4.8)から $\bar{\pi}_0(\cdot|i)$ を適当に与えれば1つの適応政策 $\bar{\pi}=(\bar{\pi}_0, \bar{\pi}_1, \cdots)$ が構成される．このとき，任意の $i \in S$，$k \in A$ に対して，$\bar{\pi}_0(k|i)>0$ かつ $\lim_{n\to\infty}\lambda_n=0$，$\lim_{n\to\infty}b_n=0$，$\sum_0^\infty b_n=\infty$ が成り立つならば，推定量 $\bar{q}(n)$ の一致性と確率1で $\lim_{n\to\infty}\bar{V}_n=u^*$，$\lim\sum_{k\in A^*[q]_i}\bar{\pi}_n(k|i)=1$ が示される．これは，$\bar{\pi}_n$ はいくらでもパラメータの値 q に対する平均最適な政策に近づくことを意味している．また π^* は平均最適な適応政策でもある([10])．

5. おわりに

適応型の問題では，未知のパラメータに関する情報の収集・処理およびシステムの最適化という2つの側面があり，両者を統一的にとり扱うのが適応制御過程といえる．いわば，統計学の諸理論とORの決定モデルの解析手法とが融合した新しい理論といえよう．したがって，適応制御的な手法はもっと現実の社会の中に浸透していってもおかしくないと思えるのだが意外にそうではない．これは，たとえば問題をDPで定式化しても，一般に状態空間が高次元の集合になり実際に解を求めるのが困難のためであろう．しかし，計算アルゴリズムの開発と高速電算機の利用により，適応型の問題解決の手法は現実の問題処理に対して有効な戦力になりつつあると思われる．

参考文献

[1] Bellman, R. E.: *Dynamic Programming*. Princeton University Press, 1957

[2] Bellman, R.E.: *Adaptive Control Processes; A Guided Tour*. Princeton Univ. 1962

[3] Federgruen, A. and Schweitzer, P. T.: Non-stationary Markov decision problems with converging parameters. *J.O.T.A.* 34 (1981)

[4] 古川・門田：マルコフ決定過程の展望，第4回数理計画シンポジウム論文集，1983，111-141

[5] ハワード，R・A著，関根他訳：ダイナミックプログラミングとマルコフ過程，培風館，1971

[6] 金子哲夫：マルコフ決定理論入門，槇書店

[7] Kurano, M.: Discrete-time MDP with an Unknown Parameter. *J. Oper. Res, Soci. Japau*, **15** (1972), 67-76

[8] Kurano, M.: Adaptive Polices in MDP with Uncertain Transition Matrices. *J. Inf. & Opti. Sci.* **4** (1983), 21-40

[9] 蔵野・安田・中神：不確実情報のMDPと応用，第4回数理計画シンポジウム論文集，1983，159-178

[10] Kurano, M.: Learning Algorithms for MDP (Preprint), 1985

[11] Lakshmivarahan, S.: *Learning Algorithms: Theory and Applications*. Springer, 1981

[12] Mandl, P.: Estimation and Control in Markov Chains. *Adv. Appl. Prob.* **6** (1974), 40-60

[13] 宮沢光一：情報・決定理論序説，岩波，1971

[14] 大野勝久：マルコフ過程の計算アルゴリズム，第4回数理計画シンポジウム論文集，1983

[15] Petkau, A. T.: Sequential Medical Trials for Comparing an Experimental with a Standard Treatment. *J.A.S.A.* 73 (1978), 328-338

[16] 佐藤他：ベイズ的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．電気電信学会論文誌，J61-D，1978，1-8

[17] 坂口　実：動的計画法，至文堂，1968

[18] 坂本武司：マルコフ決定過程，情報科学講座「マルコフ過程」共立出版，1966，106-169

[19] Van Hee, K.M.: *Bayesian Control of Markov Chains*, Mathe. Centre Tracks, Amesterdam, 1978

[20] Yakomitz, S.J.: *Mathematics of Adaptive Control Processes*. Elsevier, New York, 1969