1995年度日本オペレーションズ・リサーチ学会
春季研究発表会

計画期間が不確実なマルコフ決定過程

01012560　東京工業大学　*飯田 哲夫　IIDA Tetsuo
01601360　東京工業大学　森 雅夫　MORI Masao

1　はじめに

計画期間が確率的な場合、どのような決定をしていくのが最適だろうか。終了時点が確率的にしか分からないとき(例えば、プロジェクトの抜本的な変更の行なわれる時期が明確に分かっていないとき)は、その終了確率に依存した決定をしなければならない。そこで、本研究ではあらかじめ終了確率が与えられている場合のマルコフ決定過程(MDP with Random Horizon)について考察する。

特別な場合として、終了確率が幾何分布に従うとき、そのMDPはDiscounted MDPと同値である(Ross[4])。

MDPRHにおいては最適な定常政策は存在しない。しかし、MDPRHにおける最適方程式を導出することで、有限のサポートの場合、有限期間の問題として解ける。また、無限のサポートの場合についての最適政策を求めるアルゴリズムについても議論する。

関連する研究として、Shapiro[5]は、homogeneous discounted MDPにおいてturnpike planning horizon theoremを示しており、また、Bean and Smith[2]らは、決定的な問題におけるplanning horizonについて、Bes and Sethi[3]らは、確率的discounted problemについて議論している。

2　諸定義と問題設定

ここで考える問題は、状態数・行動数ともに有限個のnon-homogeneous MDPで$(S, A, p_t, r_t, c_t, f_t)$で定義される。

p_t : t期の推移確率行列
r_t : $S \times A \to R$の利得関数
c_t : t期のsalvage cost
u_t : $S \to A$の決定関数
$u = (u_1, u_2, \cdots)$で政策
$\prod$: uの集合
f_t : t期に終了する確率
$V_0(i,u)$: 初期状態iでstrategy uを取ったときの期待利得
$V_0^N(i,u)$: 初期状態iでstrategy uを取ったときの最初のN期間の期待利得

また、次のことを仮定する。

仮定 1 $|r(s,a)| \le R$, $|c(s,a)| \le C$

$\prod$上の距離ρを次のように定義する。

$$\rho(u,u') = \sum_{n=1}^{\infty} 2^{-n}\phi_n(u,u')$$

ただし、

$$\phi_n(u,u') = \begin{cases} 1 & u_n'(x) \neq u_n(x) \quad \exists x \in S \\ 0 & u_n'(x) = u_n(x) \quad \forall x \in S \end{cases}$$

この距離の下で、$\prod$はコンパクトであり、また、そのcylinder subsetもコンパクトである。

命題 1　計画期間の期待値が有限ならば、総期待利得は有限となる。

Remark 2　$\forall i, \forall a, r(i,a) > 0$のとき、逆が成り立つ。

ここからは、次のことを仮定する。

仮定 2 計画期間の期待値は有限である。

3　最適方程式

$v_t(i)$をt期に状態iにいるときのt期以降の最大期待利得とする。そのとき、各期に1つ吸収状態を作ることで、以下のような最適方程式を得る。

$$v_t(i) = \max_{a \in A} \left\{ b_t(i,a) + \sum_{j \in S} p_t(j|i,a) \left(\frac{1 - \sum_{k=1}^{t} f_k}{1 - \sum_{k=1}^{t-1} f_k} \right) v_{t+1}(j) \right\} \quad (1)$$

ただし、 $b_t(i,a) = \alpha_t r_t(i,a) + (1-\alpha_t) c_t(i,a)$

$$\alpha_t = \frac{1 - \sum_{k=1}^{t} f_k}{1 - \sum_{k=1}^{t-1} f_k} \text{とおく。}$$

4 最適 starategy の性質

計画期間の分布が有限のサポートをもつ場合は、上の最適方程式により最適 strategy を求めることができる。しかし、無限のサポートをもつ場合については、次の定理が言える。

定理 3 (存在性)
仮定 *1,2* の下で、*MDPRH* には 最適 *strategy* が存在する。

MDPRH には、最適な statinary deterministic strategy または、randomized strategy は存在しない。従って、求める最適 strategy は non-stationary になっており、直接求めることは困難である。

そこで、Shapiro[5] が、示した Turnpike Horizon Theorem と同様の定理が、今回の MDPRH においても成り立つことを示す。

$\prod^*$: 全ての最適 strategy u^* の集合

u^{n^*} : 最初の n 期間の決定が n 期間問題の最適 stratety と一致している strategy

$$F^\infty = \{{}_1u : u \in \prod^*\}$$
$$F^n = \{{}_1u : u = u^{n^*}\}$$

そのとき、次の定理が成り立つ。

定理 4 (Turnpike Theorem)
$\exists N, \forall n \geq N, F^n \subset F^\infty$

上の定理より、十分大きな N に対する N 期間問題を解くことにより、最適 strategy となる第 1 期の決定をすることができる。つまり、最適な rolling strategy が存在することが分かる。

次に、最適 stratety または ϵ - 最適 strategy の第 1 期の決定を求めるためのアルゴリズムについて議論する。

$$\hat{\prod}^N = \{u \in \prod : {}_1u \notin F^N\}$$

このとき、次の定理が成り立つ。

定理 5
$\forall u \in \prod$ *such that* ${}_1u \in F^{N^c}$,
$v_0^N(i) - \hat{V}_0^N(i,u) >$
$2(R+C)\sum_{n=N+1}^{\infty}\prod_{k=1}^{n}\alpha_k$
のとき、
u は最適 *strategy* とはならない。

この定理により、最適 strategy とはならない決定を取り除いていくことができる。

Remark 6 F^N が *singleton* になり、$\forall u \in \prod$, *such that* ${}_1u \in F^{N^c}$ に対して、定理 3 の条件が成り立つとき、${}_1u \in F^N$ は、最適 *strategy* の第 1 期の決定となる。

従って、次のようなアルゴリズムが考えられる。

[Algorithm]

step 1. t = 1 とする。

step 2. $\delta_t = (R+C)\sum_{n=t+1}^{\infty}\prod_{k=1}^{n}\alpha_k$ とする。

step 3. $\forall a \in A, \xi_t^a = v_0^t(i) - V_0^t(i,a)$ を計算し、もし、$\xi_t^a > 2\delta_t$ かつ F^t が singleton ならば、終了。最適 starategy の 第 1 期の決定を得る。

step 4. $\delta_t \leq \epsilon$ のとき、終了。ϵ - 最適 strategy を得る。

step 5. $t = t+1$ とし、step 2 に行く。

5 終りに

実際に、このアプローチを用いて数値実験を行なった結果を発表当日に示す予定である。

参考文献

[1] Alden,J.M. and R.Smith (1992). "Rolling horizon procedures in nonhomogeneous Markov decision processes," *Operations Research* 40 183-194.

[2] Bean,J. and R.Smith (1984). "Conditions for the existence of planning horizons," *Math. of Operations Research* 9 391-401.

[3] Bes,C. and S.Sethi (1984). "Concepts of forecast and decision horizons : applications to dynamic stochastic optimization problems," *Math. of Operations Research* 13 295-310.

[4] Ross,S.M.(1984). *Introduction to Stochastic Dynamic Programming*, Academic Press,NY.

[5] Shapiro,J.F. (1968). "Turnpike planning horizons for a Markovian decision model," *Management Sci.* 14 292-300.