1996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

制約付き非定常マルコフ決定過程

01012560　東京工業大学　飯田 哲夫　IIDA Tetsuo

1　はじめに

マルコフ決定過程 (Markov Decision Processes, MDP) において非定常な場合の研究が最近多くなされてきている。その研究の多くは、割り引きありの場合および 割り引き無しの場合両方含めて制約なしの期待利得最大化問題を取り扱っている。しかしながら、最適化を行うとき、問題に全く制約がないということは考えにくい。本研究では、非定常な場合の制約付きの平均期待利得最大化問題を取り扱う。ただし、ここで扱う制約は、割り引かれた総期待コスト $=\alpha$ とする。

用いる最適化基準 (optimality criterion) は、algorithmic optimality を用いる。また、無限期間問題の最適戦略はほとんどの場合非定常な戦略となるため、決定関数列をすべて同時に求めることは非常に困難である。従って、ここでは、Turnpike Planning Horizon Theorem が制約付き平均期待利得最大化問題においても成り立つことを示し、最適な rolling 戦略の存在を示す。

関連する研究として、Hopp, Bean and Smith[2] 制約なしの平均期待利得最大化問題を取り扱っている。Guo[1] は、非定常な場合の目的・制約ともに割り引かれた問題にについて、最適戦略の存在を示している。また White[5] も、非定常な場合も含めて割り引きありの場合について議論している。しかしながら、その中では有限期間問題が中心であり、無限期間問題についてはあまり述べられていない。

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) 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 : $S \times A \to R$ のコスト関数
$\pi = (\pi_1, \pi_2, \cdots)$: 戦略
$\prod$: π の集合

$v_0(s_0, \pi)$: 初期状態 s_0 で戦略 π を取ったときの平均期待利得
$w_0(s_0, \pi)$: 初期状態 s_0 で戦略 π を取ったときの割り引かれた総期待コスト
$v_0(s_0, \pi, N)$: 初期状態 s_0 で戦略 π を取ったときの最初の N 期間の平均期待利得
$w_0(s_0, \pi, N)$: 初期状態 s_0 で戦略 π を取ったときの最初の N 期間の割り引かれた総期待コスト

そのとき、コスト制約付き平均期待利得最大化問題は、以下のように定式化される。

$$(\mathrm{P}) \left| \begin{array}{ll} sup_{\pi \in \Pi} & v(s_0, \pi) \\ s.t. & w(s_0, \pi) = \alpha. \end{array} \right.$$

制約を満たす戦略の集合を Π_α とする。戦略 $\pi^* \in \Pi_\alpha$ が、$v(s_0, \pi^*) = \sup_{\pi \in \Pi_\alpha} v(s_0, \pi)$ を満たすならば、最適戦略という。

また、次のことを仮定する。

仮定 1 $|r_t(s,a)| \le R$, $|c_t(s,a)| \le C$, *for all* t, s, a.

3　有限期間問題と Algorithmic Optimality

N 期間問題は、無限期間問題 (P) における $v(s_0, \pi)$, $w(s_0, \pi)$ を $v(s_0, \pi, N)$,$w(s_0, \pi, N)$ に置き換えた問題とする。そして、最適戦略を $\pi^*(N)$ と表す。

定理 1 (White[4],[5])　2つの *markovian deterministic* 戦略の凸結合の中に N 期間問題の最適戦略が存在する。

ここで、Hopp,Bean and Smith[2] によって提案されている最適性の基準として以下の algorithmic optimality を定義しておく。

定義 1　有限期間問題の最適戦略列の部分列 $\{\pi^*(N_m)\}_{m=1}^{\infty}$ に関して、距離 ρ の意味で $\pi^*(N_m) \to \pi^*$ *as* $m \to \infty$ となるとき、戦略 π^* を *algorithmically* 最適戦略という。

そこで、戦略集合 $\prod^{MD} \times \prod^{MD} \times [0,1]$ 内に距離 ρ を以下のように定義しておく。

$$\rho(\pi,\pi') = |p-p'| + \frac{1}{2}\sum_{t=1}^{\infty} 2^{-t}\left(\phi_t(f^1,f^{1'}) + \phi_t(f^2,f^{2'})\right),$$

ただし、

$$\phi_t(f,f') = \begin{cases} 1 & f'_t(s) \neq f_t(s) \quad \exists s \in S \\ 0 & f'_t(s) = f_t(s) \quad \forall s \in S \end{cases}$$

この距離 ρ の下で、$\prod^{MD} \times \prod^{MD} \times [0,1]$ はコンパクトとなる。

注意 1　*algorithmically* 最適戦略は存在する。

4 最適戦略と Algorithmically 最適戦略

ここでは、前節で定義した algorithmically 最適戦略と通常の最適戦略との間の関係について議論する。

また、今回扱う割り引き無しの非定常マルコフ決定過程において重要な役割を演じる弱エルゴード性を定義しておく。

定義 2　すべての l に対して、$\tau(T_l^n(\pi)) \to 0 \quad as \quad n \to \infty$ が成り立つとき、戦略 π によって定まるマルコフ連鎖は、弱エルゴード的であるという。ただし、τ は、*proper* なエルゴード係数、$T_l^n(\pi) = P_l(\pi)\cdots P_n(\pi)$ とする。

Proper なエルゴード係数として、Hajnal 係数 $\tau_1(P) = \frac{1}{2}\sup_{l,k}\sum_{j=1}^{n}|p_{lj}-p_{kj}|$ を採用する。

今回は、基礎となる非定常マルコフ連鎖に関して、弱エルゴード的であることより少し強い以下の仮定を置く。

仮定 2　すべての $k = 1,2,\cdots,$ に対して、戦略集合 Π 上で一様に、

$$\sum_{n=k}^{\infty} \tau_1(T_k^n(\pi)) \leq A_k < \infty,$$

ただし、$\frac{A_k}{k} \to 0 \quad as \; k \to \infty.$

仮定 2 が満たされる状況としては、例えば、各推移確率行列のエルゴード係数の値が、有限個を除いて、ある $\eta < 1$ によって上から抑えられていれば満たされる。

また、今回扱う制約式が意味を持つために、以下の 2 つの仮定を置く。

仮定 3　$\beta < 1$

仮定 4　任意の n に対して、n 期間問題において、$w(s_0,\underline{\pi}(n)) \leq \alpha \leq w(s_0,\overline{\pi}(n))$ で、$w(s_0,\overline{\pi}(n)) - w(s_0,\underline{\pi}(n)) > \delta$ となる戦略 $\underline{\pi}(n),\overline{\pi}(n)$ および δ が存在する。

仮定 4は設定した制約が意味のある制約になっていることを要求しているものである。

定理 2　仮定 *1,2,3,4* の下で、*algorithmically* 最適戦略は、ϵ- 最適戦略である。

5 Turnpike Planning Horizon の存在性

非定常マルコフ決定過程においては、最適戦略は一般的に非定常な戦略になっており、直接全期間の決定を求めることは困難である。そこで、Shapiro[3] が示した Turnpike Planning Horizon Theorem と同様の定理が、今回の制約付き非定常マルコフ決定過程においても成り立つことを示す。

$\prod^*_{\alpha,\epsilon}$:　全ての ϵ- 最適戦略の集合

$F_\epsilon = \{{}_1\pi : \pi \in \prod^*_{\alpha,\epsilon}\}$:制約を満たす ϵ- 最適戦略の第 1 期における決定の集合

$F^n = \{{}_1\pi : \pi = \pi^*(n)\}$: n 期間問題の制約を満たす最適戦略の第 1 期における決定の集合

そのとき、次の定理が成り立つ。

定理 3 (Turnpike Planning Horizon Theorem) 仮定 *1,2,3,4* の下で、任意の $n \geq L$ に対して、$F^n \subset F_\epsilon$ となる L が存在する。

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、最適戦略となる第 1 期の決定を求めることができる。つまり、最適な rolling 戦略が存在することが分かる。

参考文献

[1] Guo,X. : The non-stationary Markov decision model with a constraint (in Chinese). *Acta Sci. Nat. Univ. Norm. Hunan*, 16, No.2, 107-113 (1993).

[2] Hopp,W.J.,J.C.Bean and R.Smith : A new optimality criterion for nonhomogeneous Markov decision processes. *Oper. Res.*, 35, No.6, 875-883 (1987).

[3] Shapiro,J.F. : Turnpike planning horizons for a Markovian decision model. *Management Sci.*, 14, No.5, 292-300 (1968).

[4] White,D.J. : Dynamic programming and probabilistic constraints. *Oper. Res.*, 22, 654-664 (1972).

[5] White,D.J. : Utility,probabilistic constraints, mean and variance of discounted rewards in Markov decision processes. *OR Spektrum*, 9, 13-22 (1987).