数理解析研究所講究録
981巻 1997年 155-160

制約付マルコフ決定過程：ベクトル最適化によるアプローチ

長岡工業高等専門学校　涌田和芳(Kazuyoshi WAKUTA)

1．はじめに

制約を持つ割引コスト型マルコフ決定過程を考える．2種類のコストc^0とc^1がある．各初期状態ごとに期待合計割引c^1コストについて制約が与えられて，すべての初期状態について期待合計割引c^0コストを最小にする．研究の多くは平均コスト型についてであるが，割引コスト型についてもいくつかの研究がある．Kallenberg[5]，Altmanと Shwartz[1]，Altman[2]のＬＰ法，Frid[4]，Sennott[7]のラグランジュ乗数法，Liuと Liu[6]のベクトル最適化法などである．Liuと Liu[6]は，制約付マルコフ決定過程をベクトル値マルコフ決定過程と関連付け，確定的定常政策の中から制約最適な政策を求める政策改良法を示した．しかし，確率的でシステムの履歴に依存するすべての政策が使用可能ならば，制約最適な政策は確定的定常政策とは限らない．この場合，Liuと Liu[6]のアルゴリズムは適用されない．

本論では，Liuと Liu[6]と同様に制約付マルコフ決定過程をベクトル値マルコフ決定過程と関連付ける．しかし，確率的でシステムの履歴に依存するすべての政策を考え，制約最適な政策の存在とその求め方について議論する．

2．制約付マルコフ決定過程

制約付マルコフ決定過程(ＣＭＤＰ)

$S=\{0,1,2,\ldots,N-1\}$：状態空間，$A(i)=$有限集合：行動空間

$p(j|i,a), i,j\in S, a\in A(i)$：推移確率

$c^0(i,a)$，$c^1(i,a)$：コスト関数

β $(0\leqq\beta<1)$：割引因子.

Π：すべての政策の集合，Π_S：すべての確率的定常政策の集合，Π_D：すべての確定的定常政策の集合，とおく．

$$I_\pi^0(i_0)=E_\pi\left[\sum_{n=0}^{\infty}\beta^n c^0(i_n,a_n)|i_0\right],\quad I_\pi^1(i_0)=E_\pi\left[\sum_{n=0}^{\infty}\beta^n c^1(i_n,a_n)|i_0\right],\ i_0\in S.$$

$d=(d_0,\ldots,d_{N-1})$：制約ベクトル.

$$\Delta_{i_0}=\{\pi\in\Pi|I_\pi^1(i_0)\leqq d_{i_0}\},\quad \Delta=\bigcap_{i_0\in S}\Delta_{i_0}.$$

$I_{\pi^*}^0(i_0)\leqq I_\pi^0(i_0)$，$\pi\in\Delta_{i_0}$ならば，$\pi^*$は$i_0$－制約最適であるという．すべての$i_0\in S$について$i_0$－制約最適であるとき，$\pi^*$は制約最適であるという．

3．ベクトル値マルコフ決定過程との関連

$U \subset R^2$ に対して，$e(U)=\{x \in U \mid$ ある $y \in U$ に対して $y \leqq x$ ならば $y = x\ \}$ とおく．

$c(i,a)=(c^0(i,a),c^1(i,a))$ をコスト関数にもつベクトル値マルコフ決定過程(ＶＭＤＰ)を考える．

$$I_\pi(i_0) = E_\pi\left[\sum_{n=0}^{\infty}\beta^n c(i_n,a_n)\middle| i_0\right],\ i_0 \in S.$$

$$V(i_0)=\bigcup_{\pi\in\Pi}\{I_\pi(i_0)\},\ i_0 \in S\ ,\ V_D(i_0) = \bigcup_{f\in\Pi_D}\{I_f(i_0)\},\ i_0 \in S.$$

このとき，$V(i_0) = co\,V_D(i_0),\ i_0 \in S$．すべての $i_0 \in S$ に対して $I_{\pi^*}(i_0) \in e(V(i_0))$ であるとき，π^* はＶＭＤＰで最適であるという．

ＶＭＤＰに対して，$c^\lambda(i,a)= <\lambda, c(i,a)>, \lambda \in R^2$ をコスト関数にもつスカラー値マルコフ決定過程(ＭＤＰ(λ))を考える．

$$I_\pi^\lambda(i_0) = E_\pi\left[\sum_{n=0}^{\infty}\beta^n c^\lambda(i_n,a_n)\middle| i_0\right],\ i_0 \in S.$$

π^* ： ＭＤＰ(λ)で最適 $\Leftrightarrow$ $I_{\pi^*}^\lambda(i_0) \leqq I_\pi^\lambda(i_0)$,$^\forall i_0 \in S, {}^\forall \pi \in \Pi$.

Theorem 3.1. E をＶＭＤＰで最適なすべての確定的定常政策の集合とし，E の政策を $I_f^1(i_0)$ の大きさの順に並べる：$I_{f_0}^1(i_0) \leqq I_{f_1}^1(i_0) \leqq \cdots\cdots \leqq I_{f_r}^1(i_0)$．このとき

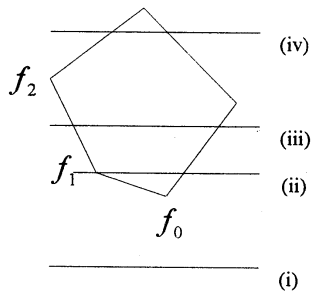
(i) $I_{f_0}^1(i_0) > d_{i_0}$ ならば，$\Delta_{i_0} = \phi$.

(ii) $I_{f_k}^1(i_0) = d_{i_0}$ ならば，f_k は i_0 －制約最適である．

(iii) $I_{f_k}^1(i_0) < d_{i_0} < I_{f_{k+1}}^1(i_0)$ ならば，i_0 －制約最適な（確率的）定常政策が存在する．

(iv) $I_{f_r}^1(i_0) \leqq d_{i_0}$ ならば，f_r は i_0 －制約最適である．

Proof. (i) (ii) (iv)は明らか．



Lemma 3.1. $\{\lambda_n\} \to \lambda$ $(\lambda_n, \lambda \in R^2)$ で，f はMDP(λ_n)，$^\forall n$ について最適とする．このとき，f はMDP(λ)で最適である．

Lemma 3.2. e_0, e_1 は, $V(i_0)$の１つの辺上の有効端点とする．このとき，e_0，e_1に対応し，共通のMDP(λ), $\lambda>0$で最適なg_0, g_1が存在する．

$f,g\in\Pi_D$, t $(0\leqq t\leqq 1)$ に対して$\pi=(t,f,g)$を，確率 t でf，確率$1-t$でgをとる政策とする．

Lemma 3.3. f, g はMDP(λ)で最適で，$I_f^1(i_0)<d_{i_0}<I_g^1(i_0)$とする．このとき，$\pi^*=(t^*,f,g)$がMDP($\lambda$)で最適で，$I_{\pi^*}^1(i_0)=d_{i_0}$となる$t^*(0\leqq t^*\leqq 1)$が存在する．

Lemma 3.4. Lemma3.3の条件を仮定する．fとgは１つの状態だけで異なると仮定する．このとき，Lemma3.3のt^*は一意に定まる．

Theorem 3.1(iii)の Proof.

$e_0=(e_0^0,e_0^1),e_1=(e_1^0,e_1^1)$を $V(i_0)$の１つの辺上の有効端点とすると $e_0^1<d_{i_0}<e_1^1$. Lemma 3.2より，e_0，e_1に対応しMDP(λ)，$\lambda>0$で最適な$g_0,g_1\in\Pi_D$が存在する．ここで，$I_{g_0}^1(i_0)<d_{i_0}<I_{g_1}^1(i_0)$.

$h_l\in\Pi_D, l=0,1,\ldots,N$

$$h_l(i)=\begin{cases}g_0(i), & i\geqq l\\ g_1(i), & i<l.\end{cases}$$

とおく．Chitgopeker[3]より，$h_l, l=0,1,\ldots,N$ は MDP(λ)，$\lambda>0$ で最適で，$I_{h_l}^1(i_0)<d_{i_0}<I_{h_{l+1}}^1(i_0)$なる$l$が存在する．Lemma 3.4より，$\pi^*=(t^*,h_l,h_{l+1})$がMDP($\lambda$), $\lambda>0$で最適で，$I_{\pi^*}^1(i_0)=d_{i_0}$となるt^*が存在する．$I_{\pi^*}^0(i_0)$は制約を満たす政策の中で最小の利得なので，$\pi^*=(t^*,h_l,h_{l+1})$はi_0-制約最適である．

Remark 3.1. Eは，ＶＭＤＰにおける政策反復によって定まる．また，Theorem 3.1(iii)の i_0－制約最適な（確率的）定常政策は，実際に求めることが可能である．したがって，制約最適な政策が存在すれば，それを求めることができる．

4．数値例

$S = \{0,1,2\}$, $A(0) = A(1) = \{0,1\}, A(2)=\{0\}$

$p(0|0,0) = 1$, $p(1|0,1) = 1$

$p(1|1,0) = 1$, $p(2|1,1) = 1$

$p(2|2,0) = 1$

$c(0,0) = (0,0)$, $c(0,1) = (-3,0)$

$c(1,0) = (2,2)$, $c(1,1) = (4,1)$

$c(2,0) = (4,1)$

$\beta = 0.5$.

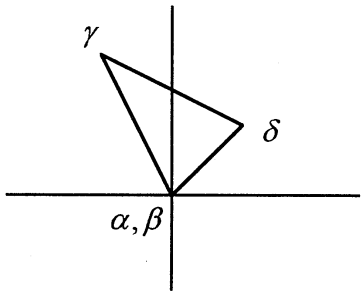
α: $\alpha(0) = 0$, $\alpha(1) = 0$, $\alpha(2) = 0$ ； β: $\beta(0) = 0$, $\beta(1) = 1, \beta(2) = 0$

γ: $\gamma(0) = 1$, $\gamma(1) = 0, \gamma(2) = 0$ ； δ: $\delta(0) = 1$, $\delta(1) = 1, \gamma(2) = 0$.

$E = \{\alpha, \beta, \gamma\}$ ：最適な確定的定常政策

$I_\alpha(0) = (0,0), I_\alpha(1) = (4,4), I_\alpha(2) = (8,2)$ ； $I_\beta(0) = (0,0), I_\beta(1) = (8,2), I_\beta(2) = (8,2)$

$I_\gamma(0) = (-1,2), I_\gamma(1) = (4,4), I_\gamma(2) = (8,2)$.



$i_0 = 0$

Eの政策を$I_f^1(0)$大きさの順に並べる.

$0 = I_\alpha^1(0) = I_\beta^1(0) < I_\gamma^1(0) = 2$.

(i) $d_0 < 0$ならば, $\Delta_0 = \phi$.

(ii) $d_0 = 0$ならば, αとβが0-制約最適.

(iv) $d_0 \geqq 2$ならば, γが0-制約最適.

(iii) $0 < d_0 < 2$のとき

$<\lambda, I_\gamma(0) - I_\alpha(0)>=0$ となる $\lambda>0$ を求める.

$I_\gamma(0)-I_\alpha(0)=(-1,2)$ なので $\lambda=(2,1)$.

E^λ を求める.

$<\lambda, I_\alpha(0)>=<\lambda, I_\beta(0)>=<\lambda, I_\gamma(0)>=0$

$<\lambda, I_\alpha(1)>=12,\ <\lambda, I_\beta(1)>=18,\ <\lambda, I_\gamma(1)>=12$

$<\lambda, I_\alpha(2)>=<\lambda, I_\beta(2)>=<\lambda, I_\gamma(2)>=18$.

$E^\lambda=\{\alpha,\gamma\}$.

E^λ の中から1つの状態だけで異なる政策のペアを求める.

α と γ は1つの状態のみで異なる.

0-制約最適な $\pi=(t,\alpha,\gamma)$ を求める.

$$I_\pi^1(i)=c^1(i,\alpha(i))+\frac{1}{2}\sum_{j\in S} p(j|i,\alpha(i))\, I_\pi^1(j),\ i=1,2\ ; \tag{1}$$

$$\begin{aligned} I_\pi^1(0)=&\, t\,c^1(0,\alpha(0))+(1-t)\,c^1(0,\gamma(0)) \\ &+\frac{1}{2}\sum_{j\in S}(t\,p(j|0,\alpha(0))+(1-t)\,p(j|0,\gamma(0)))\, I_\pi^1(j). \end{aligned} \tag{2}$$

$I_\pi^1(0)=d$ とおいて，(1) を解くと，$I_\pi^1(1)=4$, $I_\pi^1(2)=2$ ． (2)に代入して

$d=\frac{1}{2}t\,d+\frac{1}{2}(1-t)\cdot 4$． $\therefore t=(4-2d)/(4-d)$.

参考文献

[1] E.Altman and A.Shwartz, Sensitivity of constrained Markov decision processes, Ann. Oper.Res.32(1991) 1-22.

[2] E.Altman, Denumerable constrained Markov decision processes and finite approximations, Math. Oper. Res., 19(1994) 169-191

[3] S.S.Chitgopekar, Denumerable state Markovian sequential control processes :On randomizations of optimal policies, Naval Res. Logistic. Quart. 22(1975) 567-573.

[4] B.Frid, On optimal strategies in control problems with constraints, Theory Probab. Appl.17(1972)188-192.

[5] L.C.M.Kallenberg, Linear Programming and Finite Markovian Control Problems, Mathematical Centre Tracts No. 148 (Mathematisch Centrum, Amsterdam, 1983).

[6] J. Liu and K. Liu, Markov decision programming with constraints, Acta Math. Appl. Sinica, 10(1994) 1-11.

[7] L.I.Sennott, Constrained discounted Markov decision chains, Probab. Engineer. Infor. Sci., 5(1991) 463-475