数理解析研究所講究録
第1734巻 2011年 62-68

一般化割引率を伴う閾値確率問題

Threshold probability problem for a general discounted additive criterion

高知大総合人間自然科学　阪口 昌彦 (Masahiko Sakaguchi)
Graduate School of Integrated Arts and Sciences, Kochi University

1 序

動的計画法を用いるときには, 対象とする問題にしたがって最適性の原理が適用できる範囲に政策を叙述することが必要となる (cf. [4]). 確率システムを伴う動的計画問題であるマルコフ決定過程において, 加法型評価に関しては, 期待値基準を考慮する場合は各期の状態と行動を履歴として, その履歴に依存して決定をとる政策を用いる. 閾値確率基準を考慮する場合, 1つの方法としては前述の履歴に加えて各期の閾値を埋め込む定式化がある (e.g. [8, 11, 12]). この様な埋め込み法で解かれる問題は多数の研究がなされていて, 各問題によって埋め込みの仕方が工夫されている (e.g. [5, 9, 10]). Ohtsubo[9] は結合型期待値評価, また, 吉良ら [5] は各期の利得が単調に増加する成長確率基準を扱い, 各々独自のパラメータの埋め込み方法をおこなっている.

また, Iwamoto[2] により提唱された両的計画法を用いて問題を解くことも研究がなされている. [2, 3] において, 負値も認める割引率を伴う加法型期待値評価が扱われている. さらに, 埋め込み法と両的計画法の2つの方法で扱われている問題がある. 前記 [10] においては, 負値も認める乗法型期待値評価の問題が両方の方法で考えられている. 確定的システム上では, 結合型が Maruyama[6, 7] によりこれらの方法を用いて扱われた.

ここでは, 必ずシステムの状態は目標集合に到る仮定の下で, 目標集合に達するまでの総利得問題について, 負値も認める割引率を伴う加法型閾値確率評価を関係し合う2種類の埋め込みを用いて扱う. 通常, この目標集合を持つ問題はシステムの状態が目標集合に存在するとき, その集合内に留まり, かつ利得は発生しないシステムに元のシステムを再構築することにより無限期間問題に帰着させる. この問題は, 割引率を常に1かつ利得が正であるとき Ohtsubo[8] で扱われた問題となる.

2 記号と定式化

通常の閾値確率最小化問題は, 意思決定者がある値 (初期閾値)r を決定し,

$$1 \times (\text{総利得}) \leqq r$$

である確率をリスクと考え, 最小化する問題 (**最適化問題 1**) が考えられている. しかしながら, 本稿で扱う負値も認める割引率を伴う場合は次の 2 種類の最適化閾値確率問題も一般には考慮しなければならない,

$$(-1) \times (\text{総利得}) \leqq r \text{ である確率の最小化} \cdots \textbf{最適化問題 2},$$
$$0 \times (\text{総利得}) \leqq r \text{ である確率の最小化} \cdots \textbf{最適化問題 3}.$$

ここでは, 我々の問題を時間空間 $N = \{1, 2, \ldots\}$, マルコフ定常推移核 p と状態空間 S と行動空間 A 上の実数値関数 β を伴う離散時間マルコフ決定過程

$$\Gamma = ((X_n), (A_n), (Y_n), (\Lambda_n), (W_n)) \text{ on } (S, A, E, \{0, \pm 1\}, \mathbb{R})$$

として定式化する:

(i) 状態空間 S は可算,

(ii) 行動空間 $A = \bigcup_{i \in S} A(i)$ は可算, ここで $A(i)$ はシステムが状態 i にいるときの取り得る行動の集合で有限かつ空でない,

(iii) 利得空間 E は可算集合 $\{y_1, y_1, \ldots\}$, ここで各利得は $y_i \in \mathbb{R}$ $(i = 1, 2, \ldots)$ かつ E は有界, つまり, $\sup_i |y_i| < \infty$,

n 期 $(n \geqq 1)$ における状態, 行動, 利得を各々 X_n, A_n, Y_n と表記する.

(iv) 状態 i にいるときに行動 a をとるならば, システムは次のマルコフ核に従う: $i, j \in S, a \in A(i), y \in E, n \geqq 1$ に対して

$$p^a(j, y|i) = P(X_{n+1} = j, Y_n = y | X_n = i, A_n = a).$$

(v) $\beta : S \times A \to \mathbb{R}$ は割引関数で, 状態 i にいるときに行動 a をとるならば, 次の期での割引は現在の割引に割引率 $\beta(i, a)$ を乗ずることによって与えられる. ここでの割引率 $\beta(\cdot)$ は負または 0 もとり, 定数 $\beta; 0 \leqq \beta < 1$ でないことに注意する.

(vi) 意思決定者は初期閾値 $r \in \mathbb{R}$ と初期問題値 $\lambda \in \{0, \pm 1\}$ を決定する, ここで, 最適化問題 1, 2, 3 を考慮した場合 λ は各々 $1, -1, 0$ である. また, $Q = \{0, \pm 1\}$ とする. さらに, 次の問題列と閾値列を埋め込む.

1) 問題値;

$$\Lambda_1 = \lambda,$$
$$\Lambda_n = \left\{(-1) I_{(-\infty, 0)} + 0 I_{\{0\}} + 1 I_{(0, \infty)}\right\} \left(\Lambda_1 \prod_{\ell=1}^{n-1} \beta(X_\ell, A_\ell) \right),$$
$$= \left\{(-1) I_{(-\infty, 0)} + 0 I_{\{0\}} + 1 I_{(0, \infty)}\right\} \left(\Lambda_{n-1} \beta(X_{n-1}, A_{n-1}) \right),$$

2) 閾値;

$$W_1 = r,\ W_n = \begin{cases} \dfrac{W_{n-1} - \Lambda_{n-1}Y_{n-1}}{|\beta(X_{n-1}, A_{n-1})|} & \text{if } \beta(X_{n-1}, A_{n-1}) \neq 0 \\ W_{n-1} - \Lambda_{n-1}Y_{n-1} & \text{otherwise} \end{cases},\ n \geqq 2.$$

また, ある拡張した状態空間として $S_{\mathbb{R}Q} = S \times \mathbb{R} \times Q$ を用いる.

(vii) 制御の方法である政策 $\pi = (\delta_n, n \geqq 1) = (\delta_1, \delta_2, \ldots, \delta_n, \ldots)$ を次で定義する: H_n を n 期間の履歴空間とする, つまり, 各 $n \in N$ に対して, $H_1 = S_{\mathbb{R}Q}$ そして $H_{n+1} = H_n \times A \times S_{\mathbb{R}Q}$. すると, H_n はシステムが n 番目の行動を選ばなければいけない時の履歴 $h_n = (i_1, w_1, \lambda_1, a_1, i_2, w_2, \lambda_2, \ldots, a_{n-1}, i_n, w_n, \lambda_n)$ の全体の集合となる. 履歴 h_n が与えられたときの行動空間 A 上の条件付き確率 $\delta_n(a_n|h_n)$, ここで, 各 $h_n = (i_1, w_1, \lambda_1, a_1, i_2, w_2, \lambda_2, \ldots, i_n, w_n, \lambda_n) \in H_n$ に対して $\delta_n(A(i_n)|h_n) = 1$ であり, $\delta_n(a_n|\cdot)$ は H_n 上 Lebesgue 可測関数と仮定する. Δ と C を各々全ての決定ルールとその政策の集合とする. 政策 $\pi = (\delta_n, n \geqq 1)$ は任意の $n \in N$ に対して決定ルール δ_n が現在の状態 $(X_n, W_n, \Lambda_n) = (i_n, w_n, \lambda_n)$ にのみ依存した条件付き確率であるとき, マルコフと呼び, その様な決定ルールの集合を Δ_M, マルコフ政策の集合を C_M とする. また, 政策 $\pi = (\delta_n, n \geqq 1)$ は π がマルコフかつ, ある $a \in A(i)$ にその確率が集中しているとき, 確定的マルコフと呼び, $\delta_n(i, r, \lambda) = a$ と表記し, その決定ルールの集合を Δ_D, 確定的マルコフ政策の集合を C_D とする. 任意の $n \in N$ に対して $\delta_n = \delta \in \Delta_D$ のとき, $\pi = \delta^\infty$ と表記し, 定常政策と呼ぶ. そして, 定常政策の集合を C_D^s とする.

停止時刻 τ を初めて目標集合に到達する時刻とする, つまり, $\tau = \inf\{n \in N | X_n \in B\}$, ここで, そのような $n \geqq 1$ が存在しないならば $\tau = \infty$. また, **総利得**を定義する:

$$Z = \sum_{n=1}^{\tau-1} \prod_{l=0}^{n-1} \beta(X_l, A_l) Y_n,$$

ここで, $\beta(X_0, A_0) = 1$. すると, 最適化問題は次の閾値確率 $P_i^\pi(\lambda Z \leqq r)$ を, 与えられた初期閾値 r と初期問題 λ に対して, 全ての政策 π に関して最小化することになる. これらの最小化問題を簡素化するために, 次のようにマルコフ決定過程を再定義する.

Assumption 1. 目標集合 $B(\neq \phi)$ をリワードフリー, かつ, 閉じている, つまり, すべての $i \in B$, $a \in A(i)$ に対して $\sum_{j \in B} p^a(j, 0|i) = 1$.

この仮定の下では, $Z = \sum_{n=1}^{\infty} \prod_{l=0}^{n-1} \beta(X_l, A_l) Y_n$ となる. この問題の解析における都合において, 有限期間の総利得を次で定める:

$$Z_0 = 0,\quad Z_n = \sum_{k=1}^{n} \prod_{l=0}^{k-1} \beta(X_l, A_l) Y_k,\ n \geqq 1.$$

ところで, 初期状態 $X_1 = i$ と政策 π が与えられたときの事象 $\{\lambda Z \leqq r\}$ の条件付き確率を $P_i^\pi(\lambda Z \leqq r)$ と表記する. さらに, この確率過程は i だけでなく, 政策 π の取り方に依り初期閾値 r または初期問題値 λ にも依存する. したがって, 条件付き確率測度として $P^\pi_{(i,r,\lambda)}(\cdot)$ と表記するかもしれない.

以下, Assumption 1 に加えて次の仮定を伴う確率過程を考える.

Assumption 2. 全ての $\pi \in C$, 各 $(i, r, \lambda) \in S_{\mathbb{R}Q}$ に対して,

$$P^\pi_{(i,r,\lambda)}(\tau < \infty) = 1, \text{つまり}, P^\pi_{(i,r,\lambda)}(\text{ ある } n \geqq 1 \text{ に対して } X_n \in B) = 1.$$

このことは目標集合の補集合 B^c が非再帰類であることを意味する. そして, 全ての政策 $\pi \in C$, 各 $(i, r, \lambda) \in S_{\mathbb{R}Q}$ に対して, $P^\pi_{(i,r,\lambda)}(|\lambda Z| < \infty) = 1$ であることが容易にわかる.

有限·無限期間の**評価関数**と**最適値関数**を次で定める: 政策 $\pi \in C$, 各 $(i, r, \lambda) \in S_{\mathbb{R}Q}$ に対して,

$$F_n^\pi(i, r, \lambda) = P_i^\pi(\lambda Z_n \leqq r), \qquad F^\pi(i, r, \lambda) = P_i^\pi(\lambda Z \leqq r),$$
$$F_n^*(i, r, \lambda) = \inf_{\pi \in C} F_n^\pi(i, r, \lambda), \qquad F^*(i, r, \lambda) = \inf_{\pi \in C} F^\pi(i, r, \lambda).$$

次に**関数族**を定義する: ある有界区間 I に対して,

$$\mathcal{F} = \{F | F(i, \cdot, \lambda) \text{ は } \mathbb{R} \text{ 上可測かつ } F(i, r, \lambda) \in I \text{ for } i \in S, r \in \mathbb{R} \text{ and } \lambda \in Q\}.$$

$\mathcal{F}_I$ からそれ自身への**演算子** T^a, T^δ, T を定義する: $F \in \mathcal{F}_I$, $(i, r, \lambda) \in S_{\mathbb{R}Q}$, $\delta \in \Delta_M$ に対して,

$$T^a F(i, r, \lambda) = \begin{cases} \displaystyle\sum_{j \in S} \sum_{y \in E} F(j, r - \lambda y, 0) p^a(j, y | i), & \text{if } a \in A^0(i) \\ \displaystyle\sum_{j \in S} \sum_{y \in E} F\left(j, \frac{r - \lambda y}{|\beta(i, a)|}, \lambda\right) p^a(j, y | i), & \text{if } a \in A^+(i) \\ \displaystyle\sum_{j \in S} \sum_{y \in E} F\left(j, \frac{r - \lambda y}{|\beta(i, a)|}, -\lambda\right) p^a(j, y | i), & \text{if } a \in A^-(i), \end{cases}$$

$$T^\delta F(i, r, \lambda) = \sum_{a \in A(i)} T^a F(i, r, \lambda) \delta(a | i, r, \lambda),$$

$$TF(i, r, \lambda) = \inf_\delta T^\delta F(i, r, \lambda) = \min_{a \in A(i)} T^a F(i, r, \lambda),$$

ここで,

$$A^0(i) = \{a \in A(i) | \beta(i, a) = 0\},$$
$$A^+(i) = \{a \in A(i) | \beta(i, a) > 0\},$$
$$A^-(i) = \{a \in A(i) | \beta(i, a) < 0\}.$$

全ての議論において, $F,G \in \mathcal{F}_I$ に対して, $F \geq G$ は各 $(i,r,\lambda) \in S_{\mathbb{R}Q}$ に関して $F(i,r,\lambda) \geqq G(i,r,\lambda)$ を意味する.

3 最適値と最適政策

この節では無限期間において, 最適値関数が最適再帰式の一意解であることを示し, 最適定常政策の存在を得る.

先ず, 基本的な演算子の性質を与える.

Lemma 1. 有界区間 I を任意とする.

(i) $F,G \in \mathcal{F}_I$, $\delta \in \Delta$ に対して, $T^\delta F - T^\delta G = T^\delta(F-G)$.

(ii) $F,G \in \mathcal{F}_I$ かつ $F \geq G$ のとき, 各 $a \in A(\cdot)$ に対して $T^a F \geq T^a G$, 各 $\delta \in \Delta$ に対して $T^\delta F \geq T^\delta G$, かつ $TF \geq TG$.

$\pi = (\delta_n, n \geqq 1) \in C$ とある与えられた 1 期間履歴 $(i,r,\lambda,a) \in S_{\mathbb{R}Q} \times A$ に対して, ${}^1\pi^{(i,r,\lambda,a)} = (\delta_n^{(i,r,\lambda,a)}, n \geqq 1)$ を各 $h_n \in H_n$, $n \geqq 1$ について $\delta_n^{(i,r,\lambda,a)}(\cdot|h_n) = \delta_{n+1}(\cdot|(i,r,\lambda,a),h_n)$ で定義する. すると, 固定された (i,r,λ,a) に対して ${}^1\pi^{(i,r,\lambda,a)} \in C$ がわかる. 簡便さのために次の記号を用いる: $\pi = (\delta_n, n \geqq 1) \in C$, $(i,r,\lambda) \in S_{\mathbb{R}Q}$ に対して,

$$
\begin{aligned}
T^{\delta_1} F^{{}^1\pi}(i,r,\lambda) = & \sum_{a \in A^0(i)} \delta_1(a|i,r,\lambda) \sum_{j,y} F^{{}^1\pi^{(i,r,\lambda,a)}}(j, r-\lambda y, 0) p^a(j,y|i) \\
& + \sum_{a \in A^+(i)} \delta_1(a|i,r,\lambda) \sum_{j,y} F^{{}^1\pi^{(i,r,\lambda,a)}}\left(j, \frac{r-\lambda y}{|\beta(i,a)|}, \lambda\right) p^a(j,y|i) \\
& + \sum_{a \in A^-(i)} \delta_1(a|i,r,\lambda) \sum_{j,y} F^{{}^1\pi^{(i,r,\lambda,a)}}\left(j, \frac{r-\lambda y}{|\beta(i,a)|}, -\lambda\right) p^a(j,y|i).
\end{aligned}
$$

Lemma 2. $\pi = (\delta_n, n \geqq 1) \in C$ を任意とする.

(i) $\lim_{n\to\infty} F_n^\pi = F^\pi$.

(ii) 各 $n \geqq 1$ に対して, $F_{n+1}^\pi = T^{\delta_1} F_n^{{}^1\pi}$, そして $F^\pi = T^{\delta_1} F^{{}^1\pi}$. 特に, $\pi = \delta^\infty \in C_D^s$ のとき, $F^\pi = T^\delta F^\pi$.

次に, 有限期間の最適値関数の基本的な性質を与える.

Theorem 1. (i) $\{F_n^*, n \geqq 0\}$ は次の有限期間最適再帰式を満たす:

$$
F_0^* = I_{S\times[0,\infty)\times Q}, \quad F_n^* = TF_{n-1}^*, \quad n \geqq 1.
$$

(ii) 各 $n \geqq 0$ に対して, $F_n^* = F_n^\pi$ を満たす確定的マルコフ政策 $\pi \in C_D$ が存在する.

Asumption 1, 2 の下での無限期間において, 再帰式または最適再帰式が一意的に解を持つ為の重要な lemma を与える.

Lemma 3. $F, G \in \mathcal{F}_{[0,1]}$, $\delta \in \Delta_D^s$ とする. $B^c \times \mathbb{R} \times Q$ 上で $F - G \leq T^\delta(F - G)$ かつ $B \times \mathbb{R} \times Q$ 上で $F = G$ のとき, $F \leq G$.

演算子の性質と Lemma 2(ii) より次の結果を得る.

Corollary 1. $\pi \in C_D^s$ とする. F^π は $B \times \mathbb{R} \times Q$ 上 $F = I_{S\times[0,\infty)\times Q}$ を満たす**再帰式** $F = T^\delta F$ の $\mathcal{F}_{[0,1]}$ 上での一意解である.

この結果, この節における次の主定理を得る.

Theorem 2. (i) F^* は $B\times\mathbb{R}\times Q$ 上 $F = I_{S\times[0,\infty)\times Q}$ を満たす**最適再帰式** $F = TF$ の $\mathcal{F}_{[0,1]}$ 上での一意解である.

(ii) $\lim_{n\to\infty} F_n^* = F^*$.

(iii) $F^* = T^\delta F^*$ を満たす定常政策 $\pi = \delta^\infty \in C_D^s$ が存在し, π は最適である.

4 値反復法と政策改良法

Theorem 1(i) と Theorem2(ii) から, 次の値反復法が得られた:

$$F^* = \lim_{n\to\infty} T^n F_0^*, \quad F_0^* = I_{S\times[0,\infty)\times Q}.$$

Lemma 4(ii) において与えられる政策改良はよく知られている Howrad[1] における加法型期待値基準のものと類似している.

Lemma 4. $\pi = \delta^\infty \in C_D^s$ を任意とする.

(i) $F \in \mathcal{F}_{[0,1]}$ は $F \geq F^*$ かつ $B \times \mathbb{R} \times Q$ 上 $F = I_{S\times[0,\infty)\times Q}$ を満たすとする. 任意の $\delta \in \Delta_D^s$ に対して $F \leq T^\delta F$ のとき, F は最適値関数である.

(ii) $\sigma \in C_D^s$ に対して, $F^{(\delta,\sigma)} \leq F^\sigma$ のとき $F^\pi \leq F^\sigma$.

次に, 政策改良法を与える. 手順は次の通りである:

I. 初期政策 $\pi_0 = {\delta_0}^\infty \in C_D^s$ を選べ.

II. ステップ n で, 政策 $\pi_n = {\delta_n}^\infty \in C_D^s$ が与えられたとする. $F^{\pi_n} \in \mathcal{F}_{[0,1]}$ を得るために $\mathcal{F}_{[0,1]}$ 上において, $B \times \mathbb{R} \times Q$ 上 $F = I_{S\times[0,\infty)\times Q}$ を満たす方程式 $F = T^{\delta_n} F$ を解け.

III. $T^{\delta_n} F^{\pi_n} = TF^{\pi_n}$ ならば手順を止めよ. $T^{\delta_n} F^{\pi_n} \neq TF^{\pi_n}$ ならば次のステップに進め.

IV. $T^{\delta_{n+1}} F^{\pi_n} = TF^{\pi_n}$ により, 新しい改良政策 $\pi_{n+1} = {\delta_{n+1}}^\infty \in C_D^s$ を見つけよ.

V. n を $n+1$ に換えて, ステップ II に戻れ.

Corollary 5から, ステップIIにおける方程式は一意に解ける. このとき, 以下の収束定理を得る.

Theorem 3. (i) 関数列 $\{F^{\pi_n}\}$ は非増加で, F^* に収束する.
(ii) $T^{\delta_n}F^{\pi_n} = TF^{\pi_n}$ のとき, F^{π_n} は最適値関数, $\pi_n = \delta_n{}^{\infty} \in C_D^s$ は最適政策となる.

参考文献

[1] R.A. Howard, Dynamic Programming and Markov Processes. The M.I.T. Press, Massachusetts, 1960.

[2] S. Iwamoto, From dynamic programming to bynamic programming. J.Math.Anal. Appl. 177 : 56–74 (1993).

[3] S. Iwamoto, On bidecision processes. J. Math. Anal. Appl. 187 : 676–699 (1994).

[4] 岩本誠一, 20世紀の名著名論, Richard E. Bellman: Dynamic programming, 情報処理, vol.46, no.7, 2005, p.842.

[5] 吉良知文, 植野貴之, 藤田敏治, 制御マルコフ連鎖における成長確率最大化について. 不確実・不確定下での意思決定過程 (京都, 2010). 数理解析研究所講究録. 1682 : 62–69 (2010).

[6] Y. Maruyama, Associative shortest amd longest path problems. Bull. Inform Cybern. 31 : 147–163 (1999).

[7] Y. Maruyama, An invariant imbedding approach to associative shortest path problems. Math. Japonica. 50 : 469–480 (2000).

[8] Y. Ohtsubo, Optimal threshold probability in undiscounted Markov decision processes with a target set. Appl. Math. Comput. 149 : 519–532 (2004).

[9] Y. Ohtsubo, Stochastic shortest path problems with associative accumulative criteria. Appl. Math. Comput. 198 : 198–208 (2008).

[10] T. Fujita, K. Tsurusaki, Stochastic optimization of multiplicative functions with negative value. J. Oper. Res. Soc. Japan 41 no. 3 : 351–373 (1998).

[11] D.J. White, Minimizing a threshold probability in discounted Markov decision processes. J. Math. Anal. Appl. 173 : 634–646 (1993).

[12] C. Wu, Y. Lin, Minimizing risk models in Markov decision processes with policies depending on target values. J. Math. Anal. Appl. 231 : 47–67 (1999).