数理解析研究所講究録
第1726巻 2011年 84-96

負のマルコフ決定過程における二つの閾値確率最適化の方法

Methods for optimizing two threshold probabilities
in negative Markov decision processes

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
高知大理　大坪 義夫 (Yoshio Ohtsubo)
Faculty of science, Kochi University

1 序

無限期間マルコフ決定過程において, 各期における利得が正, 負, または割引を伴う場合に様々な評価基準が扱われてきた. 例えば, 通常の期待総和評価においては, 正の場合はBlackwell[2], 負の場合はStrauch[15], 割引を伴う場合はBlackwell[1]等が挙げられる. リスク鋭感的期待総和または平均評価においては, Cavazos-Cadena[3](正の場合), Jaśkiewicz[6](負の場合)がある. また, 利得の総和がある値以下である確率を扱う閾値確率評価では正の場合([10]), 割引を伴う場合(e.g. [16], [17])がある. 本稿では, ある閾値確率評価に関して利得が負の場合を考える.

ところで, ここでの意思決定者はある閾値rの下側に目標集合に到達するまでの総利得がどのくらい散らばっているかをリスクと考えている. この閾値確率最小化問題は通常次の(1)式の評価関数を考慮することが多いが, ここでは次の(2)式の評価関数も導入する,

$$F^{\pi}(i,r)=P_i^{\pi}\left(\sum_{n=1}^{\tau-1}Y_n\leq r\right), \tag{1}$$

$$G^{\pi}(i,r)=P_i^{\pi}\left(\sum_{n=1}^{\tau-1}Y_n< r\right), \tag{2}$$

ここでiは初期状態, Y_nはn期での利得, πは意思決定者がとる政策, τは目標集合に到達する時刻である. この研究では, 上記二つの最小化問題を考慮することにより, 各々の最適値関数の特徴付けを行い, 一方の最適値関数と最適政策から他方のそれらを導く方法を結果として得た.

2 宝くじのモデル

ここで, 利得が正の場合であるが次の1段階決定問題を閾値確率評価について考えてみる. 20パーセントで100円, 5パーセントで1000円がもらえる100円の宝くじを考慮する. 100円手元にあるとして, Yをくじを買う, または買わないとしたときの1期間後の利得とする. したがって, $P^{\pi}(Y \leq r)$を最小とする政策π^*とその値を求める問題となる. ところで, 各rに対してそれぞれの行動をとったときの閾値確率$P^{\pi}(Y \leq r)$は下図の様になる.

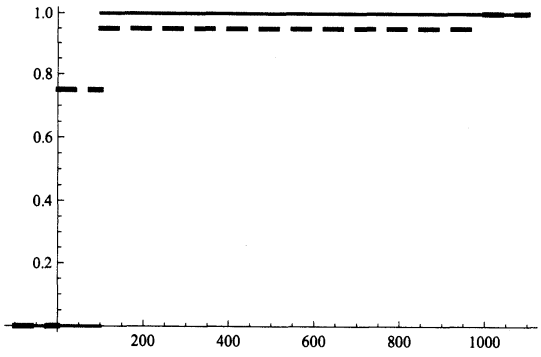

図 1: –; 宝くじを買う ––; 宝くじを買わない

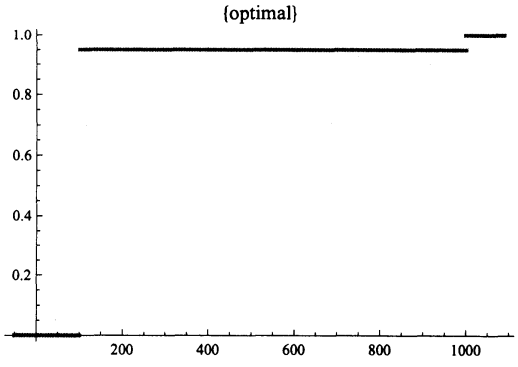

図 2: 最適な行動

ゆえに, 1つの最適な政策π^*は意思決定者が所持金100円未満の金額以下をリスクと考えているならば宝くじを買わない行動をとり, 100円以上の金額以下をリスクとするならば宝くじを買う行動をとる.

ところで, 其々の行動をとったときの期待値も考えてみると, 下記の様になる.

宝くじを買う; 期待値は$0 \times 0.75 + 100 \times 0.20 + 1000 \times 0.05 = 70$(円),

宝くじを買わない; 期待値は$100 \times 1.00 = 100$(円),

最適な行動; (擬似的な) 期待値は $100 \times 0.95 + 1000 \times 0.05 = 145$(円). ここで, 期待値最大化に関して考慮するならば, 宝くじを買わない方が買う行動をとるよりも高くなっている. このことは現実の宝くじに即している.

3 記号と定式化

この節では我々の下方リスク最小化問題を負の離散時間非割引マルコフ決定過程として定式化する:

(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 $\{y_1, y_2, \ldots\}$, ここで各利得は y_i $(i = 1, 2, \ldots)$ は非正かつ E は有界, つまり, $-\infty < \inf_i y_i < y_i \leq 0$,

n 期 $(n \geq 1)$ 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: $i, j \in S, a \in A(i), y \in E, n \geq 1$ に対して

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

また, ある拡張した状態空間として $S_R = S \times R$ を用いる, ここで $R = (-\infty, \infty)$.

目標集合 $B \subset S$ を空でないとする. 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, $\tau = \inf\{n | X_n \in B\}$, ここで, そのような $n \geq 1$ が存在しないならば $\tau = \infty$. また, 総非負利得を定義する:

$$Z = \sum_{k=1}^{\tau - 1} Y_k.$$

すると, 通常の問題は次の閾値確率 $P_i^\pi(Z \leq r)$ を全ての政策 π と与えられた初期閾値 r に関して最小化することになる. しかしながら, その

閾値確率最小化問題を直接解析するには困難な面がある. 例えば, 初期パラメータであるrに関するある種の連続性に関する問題である. したがって, 別の閾値確率$P_i^{\pi}(Z < r)$を伴うもう一つのリスク最小化問題を導入する. 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. 目標集合Bをある再帰類かつリワードフリーとする, 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, $Z = \sum_{k=1}^{\infty} Y_k$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:

$$Z_0 = 0, \quad Z_n = \sum_{k=1}^{n} Y_k, \ n \geq 1.$$

さらには, 意思決定者は各期における閾値に依存して行動をとるかもしれない. したがって, それらの閾値列を定義する:

$$W_1 = r, \quad W_n = W_1 - Z_{n-1} = W_{n-1} - Y_{n-1}, \quad n \geq 2,$$

ここでrはある与えられた初期閾値である. それゆえ, H_nをn期間の履歴空間とする: 各$n \in N$に対して, $H_1 = S_R$そして$H_{n+1} = H_n \times A \times S_R$. 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$h_n = (i_1, w_1, a_1, i_2, w_2, \ldots, a_{n-1}, i_n, w_n)$の全体の集合となる. 政策$\pi = (\delta_n, n \geq 1) = (\delta_1, \delta_2, \ldots, \delta_n, \ldots)$を次で定義する: 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$h_n = (i_1, w_1, a_1, i_2, w_2, \ldots, i_n, w_n) \in H_n$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$\pi = (\delta_n, n \geq 1)$は任意の$n \in N$に対して決定ルール$\delta_n$が現在の状態$(X_n, W_n) = (i_n, w_n)$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$\pi = (\delta_n, n \geq 1)$は$\pi$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, $\delta_n(i, r) = a$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.

初期状態 $X_1 = i$ と政策 π が与えられたときの事象 $\{Z \leq r\}$ の条件付き確率を $P_i^\pi(Z \leq r)$ と表記する. 同様に, $\{Z < r\}$ の場合も $P_{(i,r)}^\pi(Z < r)$ と表記する. ところで, 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 にも依存する. したがって, 条件付き確率測度として $P_{(i,r)}^\pi(\cdot)$ と表記するかもしれない. この報告を通して $P_{(i,r)}^\pi(\tau < \infty) = 1$ と仮定する, つまり, 全ての $\pi \in C$, 各 $(i,r) \in S_R$ に対して, $P_{(i,r)}^\pi(\text{ ある } n \geq 1 \text{ に対して } X_n \in B) = 1$. このことは目標集合 B がただ一つの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 $(i,r) \in S_R$ に対して, $P_{(i,r)}^\pi(-\infty < Z \leq 0) = 1$ であることが容易にわかる.

確定的決定ルール $\delta \in \Delta_D$ は各 $(i,r) \in S_R$ に対して, $0 \leq u < \mu$ なる全ての u に関して $\delta(i,r) = \delta(i, r-u)$ (resp. $\delta(i,r) = \delta(i, r+u)$) を満たす正数 μ が存在するならば R 上左 (resp. 右) 連続と呼ぶ. 政策 $\pi = (\delta_n, n \geq 1) \in C_D$ は各 $n \geq 1$ に対して決定ルール δ_n が左 (resp. 右) 連続ならば左 (resp. 右) 連続と呼ぶ. 我々の問題は二つの閾値確率 $P_i^\pi(Z \leq r)$, $P_i^\pi(Z < r)$ を全ての政策 $\pi \in C$ に関して最小化することである. 一番目のリスク最小化問題を $(\mathcal{P})_\leq$ と表記し, 二番目を $(\mathcal{P})_<$ と表す. 有限·無限期間の評価関数と最適値関数を次で定める: 各 $(i,r) \in S_R$, $\pi \in C$ に対して,

$$
\begin{aligned}
F_n^\pi(i,r) &= P_i^\pi(Z_n \leq r), & F^\pi(i,r) &= P_i^\pi(Z \leq r),\\
G_n^\pi(i,r) &= P_i^\pi(Z_n < r), & G^\pi(i,r) &= P_i^\pi(Z < r),\\
F_n^*(i,r) &= \inf_{\pi \in C} F_n^\pi(i,r), & F^*(i,r) &= \inf_{\pi \in C} F^\pi(i,r),\\
G_n^*(i,r) &= \inf_{\pi \in C} G_n^\pi(i,r), & G^*(i,r) &= \inf_{\pi \in C} G^\pi(i,r).
\end{aligned}
$$

次に関数族を定義する: $\mathcal{F}_I$ は各 $i \in S$ に対して $F(i,\cdot)$ は R 上可測である S_R からある有界区間 I への写像 F の集合, そして,

$$\mathcal{F}_f = \{F \in \mathcal{F}_{[0,1]} | F(i,r) = 1 \text{ for } i \in S \text{ and } r > 0\},$$
$$\mathcal{F}_g = \{F \in \mathcal{F}_{[0,1]} | F(i,r) = 1 \text{ for } i \in S \text{ and } r \geq 0\},$$
$$\mathcal{F}_\ell = \{F \in \mathcal{F}_f | i \in S \text{に対して} F(i,\cdot) \text{は単調非減少}, R \text{上左連続}\},$$
$$\mathcal{F}_r = \{F \in \mathcal{F}_g | i \in S \text{に対して} F(i,\cdot) \text{は単調非減少}, R \text{上右連続}\}.$$

$F \in \mathcal{F}_\ell$に対してF_rをFの右連続化とする, つまり, 任意の$(i,r) \in S \times R$に関して$F_r(i,r) = \lim_{s \downarrow r} F(i,s)$. 同様に, $F \in \mathcal{F}_r$に対してF_ℓをFの左連続化とする. これらの関数はwell definedで$F_r \in \mathcal{F}_r$かつ$F_\ell \in \mathcal{F}_\ell$. Lemma 6(ii)から$G^* \in \mathcal{F}_\ell$を示し, section 6において$F^* \in \mathcal{F}_r$を得る方法を提案する. $\mathcal{F}_I$からそれ自身への演算子T^a, T^δ, Tを定義する: $F \in \mathcal{F}_I$, $(i,r) \in S_R$, $a \in A(i)$, $\delta \in \Delta_M$に対して,

$$T^a F(i,r) = \sum_{j \in S} \sum_{y \in E} F(j, r-y) p^a(j,y|i),$$
$$T^\delta F(i,r) = \sum_{a \in A(i)} T^a F(i,r) \delta(a|i,r),$$
$$TF(i,r) = \inf_{\delta \in \Delta_M} T^\delta F(i,r) = \min_{a \in A(i)} T^a F(i,r).$$

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

4 最適値と最適政策

この節では二つの問題における最適値関数が各々に対応する最適方程式の一意解であることを示し, 問題$(\mathcal{P})_<$においての最適左連続定常政策の存在を与える. 後に, Theorem 5(ii)において$(\mathcal{P})_\leq$における最適左連続定常政策の存在を得る.

次に基本的なlemmasを与える. これらはLemma 2.1 and 2.2 in Ohtsubo and Toyonaga[8]とLemma 3.2 in Ohtsubo[12]において与えられた.

Lemma 1. 有界区間 I を任意とする.

(i) $F, G \in \mathcal{F}_I$ と $\delta \in \Delta$ 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 $TF \geq TG$.

(iii) $F \in \mathcal{F}_\ell$ (resp. $\in \mathcal{F}_r$) のとき, 各 $a \in A(\cdot)$ に対して $T^a F \in \mathcal{F}_\ell$ (resp. $\mathcal{F}_r$). また, T は $\mathcal{F}_I$ ($\mathcal{F}_f$, $\mathcal{F}_g$, $\mathcal{F}_\ell$ または $\mathcal{F}_r$) からそれ自身への演算子.

(iv) 各 $n \geq 0$ に対して $J_n \in \mathcal{F}_\ell$ かつ $J_n \leq J_{n+1}$ のとき, $\lim_{n\to\infty} J_n \in \mathcal{F}_\ell$.

(v) 各 $n \geq 0$ に対して $K_n \in \mathcal{F}_r$ かつ $K_n \geq K_{n+1}$ のとき, $\lim_{n\to\infty} K_n \in \mathcal{F}_r$.

Lemma 2. 各 $F \in \mathcal{F}_\ell$ *(resp.* $\mathcal{F}_r$*)* に対して, $TF = T^\delta F$ を満たす左 *(resp.* 右*)* 連続決定ルール $\delta \in \Delta_D$ が存在する.

Hernández-Lerma and Lasserre ([4], Lemma 4.2.4, p.47) において与えられている lim と min の順序交換を我々のモデルに適用する.

Lemma 3. $\{F_n\}$ を $\mathcal{F}_I$ 上の非減少列とする. 各 $(i, r) \in S \times R$ に対して, $\lim_{n\to\infty} TF_n(i, r) = T \lim_{n\to\infty} F_n(i, r)$.

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

$$T^{\delta_1} F^{{}^1\pi}(i, r) = \sum_{a \in A(i)} \delta_1(a|i, r) \sum_{j,y} F^{{}^1\pi^{(i,r,a)}}(j, r - y) p^a(j, y|i).$$

Lemma 4. $\pi = (\delta_n, n \geq 1) \in C$ を任意とする.

(i) $n \geq 0$ に対して, $F_n^\pi \leq F_{n+1}^\pi \leq \lim_{n\to\infty} F_n^\pi = F^\pi$.

(ii) $n \geq 0$ に対して, $G_n^\pi \leq G_{n+1}^\pi \leq \lim_{n\to\infty} G_n^\pi = G^\pi$.

(iii) $n \geq 0$ に対して, $F_n^\pi, G_n^\pi \in \mathcal{F}_{[0,1]}$ かつ $F^\pi, G^\pi \in \mathcal{F}_{[0,1]}$.

(iv) $n \geq 0$ に対して, $F^{\pi}_{n+1} = T^{\delta_1} F^{1\pi}_{n}$ かつ $F^{\pi} = T^{\delta_1} F^{1\pi}$. 特に, $\pi = \delta^{\infty} \in C^{s}_{D}$ のとき $F^{\pi} = T^{\delta} F^{\pi}$.

(v) $n \geq 0$ に対して, $G^{\pi}_{n+1} = T^{\delta_1} G^{1\pi}_{n}$ かつ $G^{\pi} = T^{\delta_1} G^{1\pi}$. 特に, $\pi = \delta^{\infty} \in C^{s}_{D}$ のとき $G^{\pi} = T^{\delta} G^{\pi}$.

次に, 問題 $(\mathcal{P})_{<}$ における有限·無限期間の最適値関数の基本的な性質を与える. 加えて, 問題 $(\mathcal{P})_{\leq}$ に関して $F^{*} \in \mathcal{F}_{r}$(Theorem 5) を除いて同様の結果を得る.

Theorem 1. (i) $n \geq 0$ に対して, $G^{*}_{n} \in \mathcal{F}_{\ell}$ かつ $\{G^{*}_{n}, n \geq 0\}$ は次の最適方程式を満たす:

$$G^{*}_{0} = I_{(0,\infty)}, \quad G^{*}_{n} = TG^{*}_{n-1}, \quad n \geq 1.$$

.

(ii) $n \geq 0$ に対して, $G^{*}_{n} = G^{\pi}_{n}$ を満たす左連続政策 $\pi \in C_{D}$ が存在する.

(iii) $n \geq 0$ に対して, $G^{*}_{n} \leq G^{*}_{n+1} \leq \lim_{n\to\infty} G^{*}_{n} \leq G^{*}$ かつ $\lim_{n\to\infty} G^{*}_{n} \in \mathcal{F}_{\ell}$.

Theorem 2. (i) $n \geq 0$ に対して, $F^{*}_{n} \in \mathcal{F}_{r}$ かつ $\{F^{*}_{n}, n \geq 0\}$ は次の最適方程式を満たす:

$$F^{*}_{0} = I_{[0,\infty)}, \quad F^{*}_{n} = TF^{*}_{n-1}, \quad n \geq 1.$$

.

(ii) $n \geq 0$ に対して, $F^{*}_{n} = F^{\pi}_{n}$ を満たす右連続政策 $\pi \in C_{D}$ が存在する.

(iii) $n \geq 0$ に対して, $F^{*}_{n} \leq F^{*}_{n+1} \leq \lim_{n\to\infty} F^{*}_{n} \leq F^{*}$.

最適値関数 F^{*}, G^{*} を特徴付ける重要な lemma を与える.

Lemma 5. $\pi = (\delta_n, n \geq 1) \in C$ を任意とする.

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

(ii) F^π は $B \times R$ 上 $F = I_{[0,\infty)}$ を満たす $\mathcal{F}_{[0,1]}$ 上で方程式 $F = T^\delta F$ の一意解である.

(iii) G^π は $B \times R$ 上 $F = I_{(0,\infty)}$ を満たす $\mathcal{F}_{[0,1]}$ 上で方程式 $G = T^\delta G$ の一意解である.

Lemma 6. (i) $\lim_{n\to\infty} F_n^* = F^*$.

(ii) $\lim_{n\to\infty} G_n^* = G^*$, そして $G^* \in \mathcal{F}_\ell$.

この結果, この節における次の主定理を得る.

Theorem 3. (i) F^* は $B \times R$ 上 $F = I_{[0,\infty)}$ を満たす最適方程式 $F = TF$ の $\mathcal{F}_{[0,1]}$ 上での一意解である.

(ii) G^* は $B \times R$ 上 $G = I_{(0,\infty)}$ を満たす最適方程式 $G = TG$ の $\mathcal{F}_{[0,1]}$ 上での一意解である.

(iii) $G^* = T^\delta G^*$ を満たす左連続定常政策 $\pi = \delta^\infty \in C_D^s$ が存在し, 問題 $(\mathcal{P})_<$ において π は最適である.

5 値反復法と政策改良法

Theorem 1 と Lemma 6 から, 次の値反復法が得られた:

$$F^* = \lim_{n\to\infty} T^n F_0^*, \quad F_0^* = I_{[0,\infty)},$$
$$G^* = \lim_{n\to\infty} T^n G_0^*, \quad G_0^* = I_{(0,\infty)}.$$

Lemma 8 において与えられる政策改良はよく知られている Howrad[5] によるものと類似している.

Lemma 7. (i) $F \in \mathcal{F}_{[0,1]}$ は $F \geq F^*$ かつ $B \times R$ 上 $F = I_{[0,\infty)}$ 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 は問題 $(\mathcal{P})_\leq$ で最適値関数である.

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

Lemma 8. $\pi = \delta^\infty \in C_D^s$ を任意とする.

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

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

次に, 問題 $(\mathcal{P})_<$ における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. $G^{\pi_n} \in \mathcal{F}_{[0,1]}$ を得るために $\mathcal{F}_{[0,1]}$ 上において, $B \times R$ 上 $G = I_{(0,\infty)}$ を満たす方程式 $G = T^{\delta_n} G$ を解け.

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

IV. $T^{\delta_{n+1}} G^{\pi_n} = TG^{\pi_n}$ 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.

Lemma 5(iii) からステップ II における方程式は一意に解ける. 上記の手順で $B \times R$ 上 $G = I_{[0,\infty)}$ のとき, 問題 $(\mathcal{P})_\leq$ の政策改良法となる.

このとき, 以下の収束定理を得る.

Theorem 4. (i) 関数列 $\{G^{\pi_n}\}$ は非増加で, G^* に収束する.

(ii) $T^{\delta_n} G^{\pi_n} = TG^{\pi_n}$ のとき, 問題 $(\mathcal{P})_<$ における G^{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.

6 二問題間の最適値関数と最適政策の関係

ここでは問題 $(\mathcal{P})_\leq$ における最適値関数と最適政策のある連続性について関心がある. それゆえ, Lemma 6(ii) による問題 $(\mathcal{P})_\leq$ における最適値関数 $G^*(i,\cdot)$ に関する左連続性を用いて二問題間の最適値関数と最適政策の関係を考慮する.

Lemma 9. 任意の $F \in \mathcal{F}_r$ に対して, $G \in \mathcal{F}_\ell$, 各 $a \in A(\cdot)$ に関して, $(T^a G)_r = T^a G_r$, $(T^a F)_\ell = T^a F_\ell$, $(TG)_r = TG_r$ そして $TG = (TG_r)_\ell$.

Lemma 10. 各$n \geq 0$に対して, $(G_n^*)_r \leq (G_{n+1}^*)_r \leq \lim_{n\to\infty}(G_n^*)_r \leq (G^*)_r$ かつ $(G^*)_r \in \mathcal{F}_r$.

Lemma 11. $(G^*)_r$ は $B \times R$ 上 $F = I_{[0,\infty)}$ を満たす方程式 $F = TF$ の解である.

Theorem 3(i) より, 演算子 T は $B \times R$ 上 $I_{[0,\infty)}$ を満たす, $\mathcal{F}_{[0,1]}$ において一意な不動点を持つことより次の Theorem を得る.

Theorem 5. (i) $F^* = (G^*)_r$, $G^* = (F^*)_\ell$ かつ $F^* \in \mathcal{F}_r$.
(ii) $F^* = T^\delta F^*$ を満たす右連続定常政策 $\pi_n = (\delta)^\infty \in C_D^s$ 存在する.
そして, π は問題 $(\mathcal{P})_\leq$ において最適政策となる.

任意の決定ルール $\delta \in \Delta_D$ に対して, 各 $(i,r) \in S_R$, $n \geq 1$ に関して $a_n = \delta(i, r + 1/n)$ とする. $A(i)$ は有限であるので, $\alpha \in A(i)$ と $\lim_{k\to\infty} a_{n_k} = \alpha$ を満たすような $\{a_n\}$ の部分列 $\{a_{n_k}\}$ が存在する. つまり, 任意の $n_i \geq N$ に対して $a_{n_i} = \alpha$ を満たす N が存在する. それゆえ, 決定ルール $\delta_r(i,r) = \alpha$ と定義する. また, 特別な場合としては, $\lim_{s\downarrow r} \delta(i,s)$ が存在する, つまり, $r < u \leq \mu$ を満たす任意の u に対して $\delta(i,s) = \delta(i,u)$ を満足する μ が存在するとき, $\delta_r(i,r) = \lim_{s\downarrow r} \delta(i,s)$ となる.
さらには, 政策 $\pi = \{\delta_n, n \geq 1\} \in C_D$ の右連続化政策を次で定義する: $(\pi)_r = ((\delta_n)_r, n \geq 1) \in C_D$. 同様に左連続化政策 $(\sigma)_\ell = \{(\gamma_n)_\ell, n \geq 1\}$ も定義する. すると, δ_n(resp. γ_n) が左 (resp. 右) 連続かつその極限, $\lim_{s\downarrow r} \delta_n(i,s)$ (resp. $\lim_{s\uparrow r} \gamma_n(i,s)$) が存在しているとき, 政策 $(\pi)_r$ (resp. $(\sigma)_l$) は右 (resp. 左) 連続政策となる.

Lemma 12. $G \in \mathcal{F}_\ell$ かつ $\delta \in C_D$ とする. $T^\delta G = TG$ のとき, $T^{\delta_r} G_r = TG_r$.

二問題間の最適政策の関係に関する主定理を得る.

Theorem 6. $\pi = \delta^\infty \in C_D^s$ とする. π は問題 $(\mathcal{P})_<$ における最適政策のとき, $(\pi)_r$ は問題 $(\mathcal{P})_\leq$ における最適政策となる.

参考文献

[1] D. Blackwell, Discounted dynamic programming, Ann. Math. Statist. **36**, 226–235(1965).

[2] D. Blackwell, Positive dynamic programming, In *Proc.5th Berkeley Symp. on Math.Statist.Prob. Vol.1*, University of California Press, Berkeley, 415–418 (1967).

[3] R. Cavazos-Cadena, Optimality equations and inequalities in a class of risk-sensitive average cost Markov decision chains. Math. Meth. Oper. Res. 71 : 47–84 (2010).

[4] O. Hernández-Lerma, J.B. Lasserre, Discrete-Time Markov Control Processes. Basic Optimality Criteria. Springer, New York, 1996.

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

[6] A. Jaśkiewicz, A note on negative dynamic programming for risk-sensitive control. Oper. Res. Lett. 36 : 531–534 (2008).

[7] J. Neveu, Mathmatical fundations of the calculus of probability. Holden-Day, San Francisco, 1965.

[8] Y. Ohtsubo, K. Toyonaga, Optimal policy for minimizing risk models in Markov decision processes. J. Math. Anal. Appl. 271 : 66–81 (2002).

[9] Y. Ohtsubo, Minimizing risk models in stochastic shortest path problems. Math. Meth. Oper. Res. 271 : 79–88 (2003).

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

[11] Y. Ohtsubo, K. Toyonaga, Equivalence classes for minimizing risk models in Markov decision processes. Math. Method. Oper. Res. 60 : 239–250 (2004).

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

[13] M. Sakaguchi, Y.Ohtsubo, Optimal threshold probability and policy iteration in semi-Markov decision processes, Int. J. Pure Appl. Math. 59 : 225–242 (2010).

[14] M. Sakaguchi, Y.Ohtsubo, Optimal threshold probability and expectation in Markov decision processes, preprint.

[15] R.E. Strauch, Negative dynamic programming. Ann. Math. Statist. 37 : 871–890 (1966).

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

[17] 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).