数理解析研究所講究録 1306 巻 2003 年 83-90

目標集合を持つ非割引マルコフ決定過程における最適閾値確率

Optimal threshold probability in undiscounted Markov decision processes with a target set

高知大学・理学部　大坪　義夫 (Yoshio Ohtsubo)
Faculty of Science, Kochi University

1．はじめに

目標集合を持つ非割引マルコフ決定過程における閾値確率（リスク）最小化問題を考える．この問題を再帰クラスをもつ無限期間非割引マルコフ決定過程として定式化する．主な結果として，最適値関数が最適方程式の一意な解であり，定常な最適政策が存在することを示す．また，いくつかの値反復法と政策改良法を与える．

非割引マルコフ決定過程はとても重要な最適化問題の一つであり，多くの文献で研究されている．(e.g. [2, 4, 5, 6, 14]). Eaton and Zadeh[5] はそのようなマルコフ決定過程を有限状態・有限行動（アクション）をもつ pursuit problem として定式化し，少なくとも 1 つの proper な政策が存在すれば最適政策に対応する総期待コストは最適方程式の一意な解であることを示し，値反復法によって最適値を与えた．Derman[6, 7] は 目標状態が吸収的な有限マルコフ決定を研究して first passage problem と呼んでいる．彼はその問題が最適定常政策をもつことを示し，逐次近似法，政策改良法，線形計画法を用いて最適解を求めた．Bertsekas and Tsitsiklis[1] はコストの非負性を仮定しないで確率最短路問題へと拡張した．また，Veinott[14] は [5]，[6] の結果を一時的 (transient) マルコフ決定過程へ一般化し，最適定常政策の存在を示した．さらに，Hernández-Lerma and Lasserre[9] はボレル状態空間とコンパクト行動空間をもつ一時的マルコフ決定過程へと拡張した．これらのすべてでは，評価関数は期待総非割引利得（またはコスト）$E[\sum_{k=1}^{\infty} Y_k]$ である．ここで，Y_k は時刻 k での利得を表す．

他方，多くの研究者 [3, 8, 13, 15, 17] は，政策 π に関して閾値確率 $P_i^{\pi}(Z_{\beta} \leq r)$ を最小にするリスク最小化モデルを研究している．ここで，$Z_{\beta} = \sum_{k=1}^{\infty} \beta^k Y_k$ は総割引利得，r は閾値で i は初期状態である．White[16] は有界な利得集合をもつ有限マルコフ決定過程でそのような問題を考えているが，彼の Lemma 3 は一般には成立せず，したがって最適値の一意性も最適解の存在も証明されていない．Wu and Lin[17] は有限・無限期間の最適値関数が閾値の分布関数であることを示し，有限期間での最適定常政策の存在を示している．Ohtsubo and Toyonaga[10] は無限期間での右連続な最適定常政策の存在に関する 2 つの十分条件を与えている．これらのすべては割引マルコフ決定過程に関する結果で，Ohtsubo and Toyonaga[11] で与えられた第 1 の同値類の問題に対応している．[12] で著者は，閾値確率が $P_i^{\pi}(Z > r)$ である確率最短路問題のリスク最小化を考えている．ここで，$Z = \sum_{k=1}^{\infty} Y_k$ は総非割引コストである．この問題は，[11] での第 2 の同値類に対応している．そこでは，その問題を非割引マルコフ決定過程として定式化し，最適値関数が最適方

程式の一意な解であることを示し，右連続な最適定常政策の存在を与え，値反復法を求めた．

この報告では，閾値確率 $P_i^\pi(Z \leq r)$ に関するリスク最小化問題を考える．ここで，$Z = \sum_{k=1}^{\tau-1} Y_k$ で，τ は与えられた目標集合への初期到達時間である．τ の有限性の仮定の下，最適値関数が最適方程式の一意解であることを示し，右連続な最適定常政策の存在を与える．また，値反復法と政策改良法を与える．

2．問題の定式化

離散時間 $N = \{1, 2, \ldots\}$ 上の非割引マルコフ決定過程 $\Gamma = ((X_n), (A_n), (Y_n), p)$ を次のように定義する：状態空間 S は可算集合で，時刻 $n \in N$ における状態を X_n と表す; 行動空間 $A = \cup_{i \in S} A(i)$ は可算集合で，$A(i)$ は状態が $i \in S$ のときの実行可能な有限集合とし，時刻 $n \in N$ における行動を A_n と表す; 利得空間 E は可算集合 $\{y_1, y_2, \ldots\}$ で $y_i (i = 1, 2, \ldots)$ は非負で有界とし，$Y_n \in E$ は時刻 $n \in N$ における利得を表す確率変数とする．各 $i, j \in S, a \in A(i), y \in E$ に対して，時間的一様な確率分布を

$$q^a(j|i) = P(X_{n+1} = j | X_n = i, A_n = a),$$
$$\hat{q}_{ij}^a(y) = P(Y_n = y | X_n = i, X_{n+1} = j, A_n = a)$$

と定め，

$$p^a(j, y|i) = q^a(j|i)\hat{q}_{ij}^a(y) = P(X_{n+1} = j, Y_n = y | X_n = i, A_n = a)$$

とおく．また，新状態空間を $S_R = S \times (-\infty, \infty)$ とする．

目標集合 B を S の空でない部分集合とし，停止時間 τ を $X_n \in B$ となるような最小の $n \geq 0$ とする．総非割引利得を

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

によって定義する．そのとき，最適化問題は与えられた閾値 r に対して閾値確率 $P_i^\pi(Z \leq r)$ をすべての政策 π に関して最小にすることである．この最適化問題を簡単にするために，B は再帰クラスで "reward-free"，すなわち，すべての $i, j \in B$, $a \in A(i)$ に対して，$\sum_{j \in B} q^a(j|i) = 1$, $\hat{q}_{ij}^a(0) = 1$ と仮定する．この仮定のもとでは，

$$Z = \sum_{k=1}^{\infty} Y_k$$

である．この問題を解析するために，有限期間の総非割引利得 Z_n と確率変数 W_n を次で定める：

$$Z_0 = 0, \quad Z_n = \sum_{k=1}^{n} Y_k, \ n \geq 1,$$
$$W_1 = r, \quad W_n = W_1 - Z_{n-1} = W_{n-1} - Y_{n-1}, \quad n \geq 2,$$

政策 $\pi = (\delta_n, n \geq 1) = (\delta_1, \delta_2, \ldots, \delta_n, \ldots)$ を次のように定義する：履歴 $h_n = (i_1, w_1, a_1, i_2, w_2, \ldots, a_{n-1}, i_n, w_n)$ を $\theta_n = (X_1, W_1, A_1, X_2, W_2, \ldots, A_{n-1}, X_n, W_n)$ の実現値とする．時刻 n までの履歴の全体を H_n とする．δ_n は条件付確率 $\delta_n(a_n|h_n) = P(A_n = a_n|\theta_n = h_n)$ で与えられ，$h_n \in H_n$ について Lebesgue 可測とする．このような政策 π の全体を C とする．政策 $\pi = (\delta_n, n \geq 1)$ が Markov とは δ_n が $(X_n, W_n) = (i_n, w_n)$ のみの関数のときをいい，確定的定常とは π が Markov で，すべての $n \in N$ で $\delta_n = \delta_{n+1}$ をみたし，δ_n が確定的に $a_n \in A$ を決定するときをいう．それぞれの全体を C_M, C_D とする．$\pi = (\delta, \delta, \ldots, \delta, \ldots) \in C_D$ のとき，$\pi = \delta^\infty$ とかき，$\delta(a|i,r) = 1$ ならば $\delta(i,r) = a$ とかく．それぞれに対応する決定ルール δ の全体を $\Delta, \Delta_M, \Delta_D$ とする．

初期状態 $X_1 = i$ と政策 π を与えたときの事象 $\{Z \leq r\}$ の条件付確率を $P_i^\pi(Z \leq r)$ と表す．過程は i, π, r に依存するので，記号として条件付確率 $P_{(i,r)}^\pi(\cdot)$ を用いることもある．この報告を通して，すべての政策 $\pi \in C$ と各 $(i,r) \in S_R$ に対して，$P_{(i,r)}^\pi(\tau < \infty) = 1$ と仮定する．すなわち，$P_{(i,r)}^\pi(X_n \in B \text{ for some } n \geq 1) = 1$ であり，これは B^c が一時クラスである．このとき，$P_{(i,r)}^\pi(Z < \infty) = 1$ である．

決定ルール $\delta \in \Delta_D$ が右連続とは，各 $(i,r) \in S_R$ に対して正の実数 μ が存在して，すべての $u : 0 \leq u < \mu$ 対して $\delta(i,r) = \delta(i, r+u)$ となることである．また，政策 $\pi = (\delta_n, n \geq 1) \in C_D$ が右連続とは，各 $n \geq 1$ に対して δ_n が右連続のことである．

有限・無限期間の評価関数を

$$F_n^\pi(i,r) = P_i^\pi(Z_n \leq r), \quad F^\pi(i,r) = P_i^\pi(Z \leq r), \quad (i,r) \in S_R, \quad \pi \in C$$

とし，最適値関数を次で定める：

$$F_n^*(i,r) = \inf_{\pi \in C} F_n^\pi(i,r), \quad F^*(i,r) = \inf_{\pi \in C} F^\pi(i,r), \quad (i,r) \in S_R.$$

$F^*(i,r) = F^\pi(i,r)$, $(i,r) \in S_R$ をみたすとき，政策 $\pi \in C$ は最適であるという．

関数空間 $\mathcal{F}_r$ を，S_R から $[0,1]$ への関数 F で次をみたすものの全体とする：$i \in S$ に対して，$r < 0$ のとき $F(i,r) = 0$ で，$F(i,\cdot)$ は非減少で右連続である．一般には，$F^\pi \notin \mathcal{F}_r$ である (cf. 例 5.1)．

作用素 T^a, T^δ, T を次で定める：$F \in \mathcal{F}_r$, $(i,r) \in S_R$, $a \in A(i)$, $\delta \in \Delta$ に対して，

$$T^a F(i,r) = \int_{S_R} F(j, r-y) dp^a(j,y|i),$$
$$T^\delta F(i,r) = \sum_{a \in A(i)} T^a F(i,r) \delta(a|i,r),$$
$$TF(i,r) = \inf_\delta T^\delta F(i,r) = \min_{a \in A(i)} T^a F(i,r).$$

3．最適値と最適政策

最初にいくつかの基本的な補題を与える．

補題 3.1. (i) $F,G\in\mathcal{F}_r,\delta\in\Delta$ に対して，$T^\delta F-T^\delta G=T^\delta(F-G)$.

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

(iii) $G\in\mathcal{F}_r$ のとき，各 $a\in A(\cdot)$ に対して，$T^aG\in\mathcal{F}_r$ かつ $TG\in\mathcal{F}_r$.

補題 3.2. 各 $F\in\mathcal{F}_r$ に対して，$TF=T^\delta F$ をみたす右連続な決定ルール $\delta\in\Delta_D$ が存在する．

政策 $\pi=(\delta_n,n\geq1)\in C$ と履歴 $(i,r,a)\in S_R\times A$ に対して，政策 ${}^1\pi^{(i,r,a)}=(\delta_n^{(i,r,a)},n\geq1)$ を $\delta_n^{(i,r,a)}(\cdot|h_n)=\delta_{n+1}(\cdot|(i,r,a),h_n)$，$h_n\in H_n$，$n\geq1$ によって定義する．そのとき，固定した (i,r,a) に対して，${}^1\pi^{(i,r,a)}\in C$ である．簡単さのために，$\pi=(\delta_n,n\geq1)\in C$，$(i,r)\in S_R$ に対して，記号

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

を用いる．

補題 3.3. $\pi=(\delta_n,n\geq1)\in C$ を任意とする．

(i) 各 $n\geq0$ に対して，$F_n^\pi\geq F_{n+1}^\pi\geq\lim_{n\to\infty}F_n^\pi=F^\pi$.

(ii) 各 $n\geq0$ に対して，$F_n^\pi\in\mathcal{F}_r$ かつ $F^\pi\in\mathcal{F}_r$.

(iii) 各 $n\geq0$ に対して，$F_{n+1}^\pi=T^{\delta_1}F_n^{{}^1\pi}$ かつ $F^\pi=T^{\delta_1}F^{{}^1\pi}$. 特に，$\pi=\delta^\infty\in C_D^s$ のとき，$F^\pi=T^\delta F^\pi$.

そこで，有限・無限期間の最適値関数の基本的な性質を与える．

定理 3.1. (i) 各 $n\geq0$ に対して，$F_n^*\in\mathcal{F}_r$ かつ $\{F_n^*,n\geq0\}$ は次の最適方程式をみたす：

$$F_0^*=I_{[0,\infty)},\ F_n^*=TF_{n-1}^*,n\geq1.$$

(ii) 各 $n\geq0$ に対して，$F_n^*=F_n^\pi$ となる右連続な政策 $\pi\in C_D$ が存在する．

(iii) 各 $n\geq0$ に対して，$F_n^*\geq F_{n+1}^*\geq\lim_{n\to\infty}F_n^*=F^*$ かつ $F^*\in\mathcal{F}_r$.

定理 3.1 から，$F^*=\lim_{n\to\infty}T^nF_0^*$ であることがわかる．最適値 F^* を特徴付けるために，次の重要な補題を必要とする．

補題 3.4. $\pi=\delta^\infty\in C_D$ を任意とする．

(i) $F,G\in\mathcal{F}_r$ 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 $B\times R$ 上で $F=I_{[0,\infty)}$ をみたす方程式 $F=T^\delta F$ の一意な解である．

この結果，次の主定理を得る．

定理 3.2. (i) F^* は $B \times R$ 上で $F = I_{[0,\infty)}$ をみたす最適方程式 $F = TF$ の $\mathcal{F}_r$ 上での一意な解である．

(ii) $B^c \times R$ 上で $F^* = T^\delta F^*$ をみたす右連続な定常政策 $\pi = \delta^\infty \in C_D$ が存在し，π は最適である．

4．値反復法と政策改良法

定理 3.1 から，一つ目の値反復法が得られた：

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

そこで，他の値反復法を与える．

定理 4.1. G を $G(i,r) = 0, i \in S, r < 0$ と $G \geq F^*$ をみたす可測な関数とする．このとき，$\{T^n G\}$ は収束し，$\lim_{n\to\infty} T^n G = F^*$.

系 4.1. 政策 $\pi \in C$ に対して，$\lim_{n\to\infty} T^n F^\pi = F^*$.

次に，政策改良法を与える．その手順は次の通りである：

1. 初期政策 $\pi_0 = (\delta_0)^\infty \in C_D$ を選べ．
2. ステップ n で，政策 $\pi_n = (\delta_n)^\infty \in C_D$ が与えられているとする．F^{π_n} を求めるために，$B \times R$ 上で $F = I_{[0,\infty)}$ をみたしている方程式 $F = T^{\delta_n} F$ を解け．
3. もし $T^{\delta_n} F^{\pi_n} = TF^{\pi_n}$ ならばこの手順をストップせよ．もし $T^{\delta_n} F^{\pi_n} \neq TF^{\pi_n}$ ならば次のステップへ進め．
4. $T^{\delta_{n+1}} F^{\pi_n} = TF^{\pi_n}$ によって，新政策 $\pi_{n+1} = (\delta_{n+1})^\infty \in C_D$ を見つけよ．
5. n を $n+1$ に換えて，ステップ (ii) へ戻れ．

このとき次の結果を得る．

定理 4.2. (i) 関数列 $\{F^{\pi_n}\}$ は非増加で，F^* に収束する．

(ii) もし $T^{\delta_n} F^{\pi_n} = TF^{\pi_n}$ ならば，F^{π_n} は最適値で，$\pi_n = (\delta_n)^\infty \in C_D$ は最適政策である．

5．数値例

最初に，$F^\pi \notin \mathcal{F}_r$ であるような政策 $\pi \in C$ を与える．

例 5.1. $S = \{1,2\}$ を状態空間とし，$\{2\}$ を目標集合とする．状態 2 は再帰的（吸収的）で reward-free とする．また，$A = \{a_1, a_2\}$ を行動空間とする．確率分布を

$$p^{a_1}(2,2|1) = 1, \quad p^{a_2}(1,1|1) = p^{a_2}(2,1|1) = 1/2.$$

で与える．政策 $\pi = \delta^\infty \in C_D$ を

$$\delta(1,r) = \begin{cases} a_1 & (r \leq 2) \\ a_2 & (\text{その他}) \end{cases}$$

と定義する．このとき，

$$F^{\pi}(1,r)=\begin{cases}0 & \text{if } r<2\\ 1 & \text{if } r=2\\ \frac{1}{2} & \text{if } 2<r<3\end{cases}$$

したがって，$F^{\pi}\notin\mathcal{F}_r$ である．

次の例では，値反復法と政策改良法で最適値と最適政策を求める．

例 5.2. $S=\{1,2,3\}$ を状態空間とし，$\{3\}$ を目標集合とする．状態 3 は吸収的で reward-free とする．また，$A=\{a_1,a_2\}$ を行動空間とする．確率分布は

$$\begin{aligned}&p^{a_1}(2,3|1)=p^{a_1}(3,2|2)=1,\\ &p^{a_2}(2,4|1)=2/3,\quad p^{a_2}(3,4|1)=1/3,\\ &p^{a_2}(2,2|2)=p^{a_2}(3,1|2)=1/2.\end{aligned}$$

である．

まず，値反復法：$F_n^*=TF_{n-1}^*$, $n\geq 1$, $F_0^*=I_{[0,\infty)}$ によって最適値を求めてみよう．明らかに，

$$F_n^*(3,r)=I_{[0,\infty)}(r),\quad r\in R,\quad n\geq 0$$

である．また，簡単に，

$$\begin{aligned}&F_1^*(2,r)=I_{[2,\infty)}(r),\quad F_1^*(1,r)=I_{[4,\infty)}(r),\\ &F_2^*(2,r)=1/2I_{[2,4)}(r)+I_{[4,\infty)}(r),\\ &F_2^*(1,r)=1/3I_{[5,6)}(r)+I_{[6,\infty)}(r).\end{aligned}$$

である．帰納法により，$n\geq 3$ に対して，

$$\begin{aligned}F_n^*(2,r) &= \sum_{k=1}^{n-1}\sum_{i=1}^{k}(1/2)^i I_{[2k,2(k+1))}(r)+I_{[2n,\infty)}(r),\\ F_n^*(1,r) &= 1/3I_{[5,6)}(r)+\sum_{k=1}^{n-2}\sum_{i=1}^{k}(1/2)^i I_{[2k+4,2k+5)}(r)\\ &\quad+\sum_{k=1}^{n-2}(1/3+2/3\sum_{i=1}^{k}(1/2)^i)I_{[2k+5,2k+6)}(r)+I_{[2n+2,\infty)}(r).\end{aligned}$$

$F^*=\lim_{n\to\infty}F_n^*$ であるから，最適値は

$$\begin{aligned}F^*(3,r) &= I_{[0,\infty)}(r)\\ F^*(2,r) &= \sum_{k=1}^{\infty}(1-(1/2)^k)I_{[2k,2(k+1))}(r),\\ F^*(1,r) &= 1/3I_{[5,6)}(r)+\sum_{k=1}^{\infty}(1-(1/2)^k)I_{[2k+4,2k+5)}(r)\\ &\quad+\sum_{k=1}^{\infty}(1-2/3(1/2)^k)I_{[2k+5,2k+6)}(r).\end{aligned}$$

次に，政策改良法を考える．初期政策 $\pi_0 = (\delta_0)^\infty \in C_D$ を

$$\delta_0(i,r) = a_1,\ (i,r) \in S_R$$

とする．$F(3,r) = I_{[0,\infty)}(r)$ のもとで方程式 $F = T^{\delta_0}F$ を解くと，

$$F^{\pi_0}(2,r) = I_{[2,\infty)}(r),\quad F^{\pi_0}(1,r) = I_{[5,\infty)}(r).$$

このとき，

$$TF^{\pi_0}(2,r) = 1/2I_{[2,4)}(r) + I_{[4,\infty)}(r),$$
$$TF^{\pi_0}(1,r) = 1/3I_{[5,6)}(r) + I_{[6,\infty)}(r).$$

であるから，$T^{\delta_0}F^{\pi_0} \neq TF^{\pi_0}$ であることがわかる．$T^{\delta_1}F^{\pi_0} = TF^{\pi_0}$ を用いて，新政策 $\pi_1 = (\delta_1)^\infty \in C_D$ を求めると，

$$\delta_1(3,r) = a_1$$
$$\delta_1(2,r) = a_1 I_{(-\infty,2)}(r) + a_2 I_{[2,\infty)}(r),$$
$$\delta_1(1,r) = a_1 I_{(-\infty,5)}(r) + a_2 I_{[5,\infty)}(r).$$

となる．再び $F = T^{\delta_1}F$ を解くと，F^{π_1} は

$$F^{\pi_1}(2,r) = \sum_{k=1}^{\infty}\sum_{i=1}^{k}(1/2)^i I_{[2k,2(k+1))}(r),$$

$$F^{\pi_1}(1,r) = 1/3I_{[5,6)}(r) + \sum_{k=1}^{\infty}(1/3 + 2/3\sum_{i=1}^{k}(1/2)^i)I_{[2k+4,2k+6)}(r).$$

となる．このとき，$T^{\delta_1}F^{\pi_1}(2,r) = TF^{\pi_1}(2,r)$，$T^{\delta_1}F^{\pi_1}(1,r) \neq TF^{\pi_1}(1,r)$ となることがわかる．ここで，

$$TF^{\pi_1}(1,r) = 1/3I_{[5,6)}(r) + \sum_{k=1}^{\infty}\sum_{i=1}^{k}(1/2)^i I_{[2k+4,2k+5)}(r)$$
$$+ \sum_{k=1}^{\infty}(1/3 + 2/3\sum_{i=1}^{k}(1/2)^i)I_{[2k+5,2k+6)}(r).$$

再び $T^{\delta_2}F^{\pi_1} = TF^{\pi_1}$ を用いて，新政策 $\pi_2 = (\delta_2)^\infty \in C_D$ が

$$\delta_2(3,r) = a_1,\ \delta_2(2,r) = \delta_1(2,r),$$
$$\delta_2(1,r) = a_1(I_{(-\infty,5)}(r) + \sum_{k=1}^{\infty} I_{[2k+4,2k+5)}(r))$$
$$+a_2(I_{[5,6)}(r) + \sum_{k=1}^{\infty} I_{[2k+5,2k+6)}(r)).$$

と得られる．$F = T^{\delta_2}F$ を解くと，$F^{\pi_2}(2,r) = F^{\pi_1}(2,r)$，$F^{\pi_2}(1,r) = TF^{\pi_1}(1,r)$ を得る．よって，$T^{\delta_2}F^{\pi_2} = TF^{\pi_2}$ である．したがって，政策改良手順を終了し，最適値 F^{π_2} と最適政策 $\pi_2 = (\delta_2)^\infty$ を得る．

参考文献

[1] D.P. Bertsekas, J.N. Tsitsiklis, An analysis of stochastic shortest path problems. Math. Oper. Res. 16 : 580–595 (1991).

[2] D. Blackwell, Discrete dynamic programming. Ann. Math. Statist. 33 : 719–726 (1962).

[3] M. Bouakiz, Y. Kebir, Target-level criterion in Markov decision processes. J. Optim. Theory Appl. 86 : 1–15 (1995).

[4] E. Denardo, Contraction mappings in the theory underlying dynamic programming. SIAM Rev. 9 : 165–177 (1967).

[5] J.H. Eaton, L.A. Zadeh, Optimal pursuit strategies in discrete-state probabilistic systems. Trans. ASME Ser. D, J. Basic Eng. 84 : 23–29 (1962).

[6] C. Derman, On sequential decisions and Markov chains. Manage. Sci. 9 : 16–24 (1962).

[7] C. Derman, Finite State Markovian Decision Processes. Academic Press, New York, 1970.

[8] J.A. Filar, D. Krass, K.W. Ross, Percentile performance criteria for limiting average Markov decision processes. IEEE Trans. Automat. Control 40 : 2–10 (1995).

[9] O. Hernández-Lerma, J.B. Lasserre, Further Topics on Discrete-Time Markov Control Processes. Springer, New York, 1999.

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

[11] Y. Ohtsubo, K. Toyonaga, Equivalence classes for minimizing risk models in Markov decision processes. preprint.

[12] Y. Ohtsubo, Minimizing risk models in stochastic shortest path problems. Math. Meth. Oper. Res. 57 : (2003), to appear.

[13] M.J. Sobel, The variance of discounted Markov decision processes. J. Appl. Prob. 19 : 794–802 (1982).

[14] A.F. Veinott, Jr., Discrete dynamic programming with sensitive discount optimality criteria. Ann. Math. Statist. 40 : 1635–1660 (1969).

[15] D.J. White, Markov Decision Processes. Wiley, New York, 1993.

[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).