不確実性下でのベイズ推定マルコフ決定過程

堀口　正之

本稿では，確率過程での各行動選択の結果による状態観測からその推移法則を推定し，多段決定過程としてマルコフ決定過程を構成する．野球とヘルスケアにみられるデータ分析の実例を挙げながら，マルコフ決定過程として扱われる数理モデルとその最適化手法を説明する．

キーワード：マルコフ決定過程，推移法則未知，区間ベイズ推定，マルコフモデルとヘルスケア

1. はじめに

時刻 t でのシステム状態の観測値を $X_t(\omega)$ とする確率過程 $\{X_t\}$ を考える．標本 $\omega \in \Omega$ に対する観測値をそれぞれ $X_t(\omega) = x_t$ $(t = 0, 1, 2, 3, \ldots)$ とするとき，時刻 t でのシステムの状態 x_t に応じて，行動（意思決定）$a_t \in A(t)$ を選択する．各時刻 t での行動は，$f_t : S \to A, f_t(x_t) = a_t$ と表される．状態と行動の履歴 $h_t = (x_0, a_0, x_1, a_1, \ldots, x_t)$ に対して，次の推移法則（マルコフ性）を満たしている決定過程は，マルコフ決定過程 (Markov Decision Processes, MDPs) と呼ばれる：

$$P(X_{t+1} = x_{t+1}|h_t, a_t) = P(x_{t+1}|x_t, a_t) \qquad t = 0, 1, 2, 3, \ldots \quad (1)$$

また，任意の時刻での状態 $x \in S$ と行動 $a \in A$ によって生じる利得関数 $r(x, a)$ またはコスト関数 $c(x, a)$ を考えて，適切な評価基準での評価関数値最大化または最小化の最適化問題として，確率過程での一連の行動選択（政策）$\pi = (f_0, f_1, f_2, \ldots) \in \Pi$ の最適解（最適政策）$\pi^* = (f_0^*, f_1^*, f_2^*, \ldots)$ を見つけ出すことが，マルコフ決定過程での一つの目的である．具体的には，状態空間を $S = \{1, 2, \ldots\}$，決定空間（行動空間とも呼ぶ）を $A = \{1, 2, \ldots\}$，状態 $x_t \in S$ と決定 $a_t \in A$ が所与のときの $x_{t+1} \in S$ の確率分布を推移法則 $Q = (q(x_{t+1}|x_t, a_t))$ とおき，関数 $r : S \times A \to \mathbb{R}$ と $c : S \times A \to \mathbb{R}$ はそれぞれ，時刻 t での $x_t \in S, a_t \in A$ に対して生じる利得 $r(x_t, a_t)$ とコスト $c(x_t, a_t)$ を表すとき，$\{S, A, Q, r$（または $c)\}$ の四つの構成要素でマルコフ決定過程が定まる．評価基準として，たとえば，割引総期待利得の最大化問題

$$\text{Maximize} \quad E_x^\pi\left[\sum_{t=0}^{\infty} \beta r(x_t, a_t)\right], \quad \pi \in \Pi$$

の最適政策（最適解）$\pi^* = (f_0^*, f_1^*, \ldots) \in \Pi$ を求める．ただし，上記の期待値 E_x^π の x は初期状態を表す．

マルコフ決定過程に関する研究は，オペレーションズ・リサーチ (OR) の研究とともに進展し続けている．半世紀ほど前に，機関誌 “オペレーションズ・リサーチ” の 1968 年 2, 3, 6 月号に「数学講座 マルコフ型逐次決定過程 (1), (2), (3)」の連載がある [1]．ここ 10 年程度の間で，在庫管理のサプライチェーン分析の基礎として，マルコフ決定過程でのモデル化が大野勝久著 [2, 3] にあり，近年，マルコフ決定過程に関する最適化手法をまとめている「マルコフ決定過程―理論とアルゴリズム―」[4] が OR 分野の研究好適書である．近年の MDPs の実例の取り組みは，たとえば，Boucherie and Dijk [5] に，ヘルスケアに関する研究事例をはじめとして，輸送，生産，通信，ファイナンシャル・モデリングの諸分野の応用解析事例が抽象一般理論とともにまとめられている．また，抽象理論の第一線の研究書籍として Costa and Dufour [6] や Piunovskiy and Zhang [7] が挙げられる．これから MDPs を学ぼうとする若き研究者，学生には，これらの先端研究にも興味と関心をもたれると良い．

本稿では，学部学生あるいは大学院生にも MDPs について興味をもっていただけそうな話題をベースにしつつ，最適性，最適解を得るための方法（アルゴリズム）と分析例を紹介する．

2. 区間型マルコフ決定過程 (Controlled Markov-Set Chains)

MDPs の構成要素の推移確率行列 Q が区間表現されたもの，利得やコスト関数についても区間表現されたものを区間型マルコフ決定過程と呼ぶ．状態の個数

ほりぐち　まさゆき
神奈川大学理学部数理・物理学科
〒 259–1293 神奈川県平塚市土屋 2946
horiguchi@kanagawa-u.ac.jp

が n，すなわち，$S=\{1,\ldots,n\}$ の場合を例に，非負値行列 $L=(l_{ij}), U=(u_{ij})$，$i,j\in S$ によって，

$$\langle L,U\rangle := \big\{Q=(q_{ij})\in\mathbb{R}_+^{n\times n} | l_{ij}\leqq q_{ij}\leqq u_{ij}, \quad q_{ij}\geqq 0, \textstyle\sum_{j\in S} q_{ij}=1, i,j\in S\big\} \quad (2)$$

が推移確率行列 (cf. Hartfiel [8]) の構成要素であるマルコフ決定過程である．利得関数あるいはコスト関数については，状態 $i\in S$ で決定 $a\in A$ を選択して生じる利得 $r(i,a)$ では $r(i,a)=[\underline{r}(i,a),\overline{r}(i,a)]$ のように，左端値 $\underline{r}(i,a)$ と右端値 $\overline{r}(i,a)$ での閉区間（閉凸集合）と表現されるものや，p 次元のベクトル集合値 $r(i,a)\in\mathcal{C}(\mathbb{R}^p)$ が扱われる．ただし，$\mathcal{C}(\mathbb{R}^p)$ は p 次元実数空間 $\mathbb{R}^p$ での凸かつコンパクトであるすべての部分集合の全体を表す．n 次元の非負値有界閉区間 $\mathbb{R}_+^n\supset D_1,D_2$ に対し，ハウスドルフ距離 ρ: $\rho(D_1,D_2)=\max\{\sup_{x\in D_1}\inf_{y\in D_2}\|x-y\|,\sup_{y\in D_2}\inf_{x\in D_1}\|x-y\|\}$ が用いられる．ただし，$\|\cdot\|$ は $\mathbb{R}^n$ のユークリッド距離である．$0<\beta<1$ を割引率とする割引総期待利得を考えるとき，

$$\phi(\pi|\{Q\})=\sum_{t=0}^{\infty}\beta Q(f_1)Q(f_2)\cdots Q(f_t)r(f_t) \quad (3)$$

を評価関数とする区間型最適化モデルが構成できる．ただし，$\pi\in\Pi$ は，$f_t: S\to A, t=0,1,2,\ldots$, $\pi=(f_0,f_1,f_2,\ldots)\in\Pi$ であり，各時刻での意思決定列 $\{f_t\}$ を表す．$f: S\to A$ によって，$f_t=f, t=0,1,2,\ldots$ であるとき，π は $\pi=(f,f,f,\ldots)=f^{(\infty)}$ の定常政策を表し，この政策 $\pi=f^{(\infty)}$ を単に f と改めて表す．また，定常政策 f に対する式 (3) の評価関数値を $\phi(f^{(\infty)}|\{Q\})=\phi(f|Q)$ と表す．$Q(f_t)$ は，時刻 t での決定 f_t による区間型推移確率行列の一つを表す．たとえば，$Q(f_t)=(q(j|i,f_t(i)))\in\langle L_t,U_t\rangle$ であり，$\mathcal{Q}_t=\langle L_t,U_t\rangle, t=0,1,2,\ldots$ に対して構成される区間型マルコフ決定過程 $\{S,A,\mathcal{Q},r$（または $c)\}$ のもとで，定常政策に最適政策が存在し，また，定常政策 $f^{(\infty)}=(f,f,f,\ldots)=f$ に対して，$\mathcal{Q}_t=\mathcal{Q}(f):=\mathcal{Q}=\langle\underline{Q},\overline{Q}\rangle, t=0,1,\ldots$ とおくとき $\phi(f|\mathcal{Q})=\big\{\phi(f|Q)\big|Q\in\mathcal{Q}\big\}\in C(\mathbb{R}^n)$ が成り立つことが示される．さらに，利得関数での評価モデルのもと，定常政策 f に関する利得関数を $r(f)$ とおき，写像 $\mathcal{L}: C(\mathbb{R}_+^n)\to\mathcal{L}: C(\mathbb{R}_+^n)$ を $\mathcal{L}(f)\boldsymbol{v}=r(f)+\beta\mathcal{Q}(f)\boldsymbol{v}$，$\boldsymbol{v}\in\mathcal{C}(\mathbb{R}_+^n)$ とすると，$\mathcal{L}$ は単調増大かつ縮小写像であって，$\mathcal{L}(f)^l\boldsymbol{v}\to\phi(f|\mathcal{Q})$ $(l\to\infty)$，つまり，$\mathcal{L}(f)$ の不動点（集合値）が $\phi(f|\mathcal{Q})$ であることも示される．また，$\phi(f|\mathcal{Q})=\big[\underline{\phi}(f),\overline{\phi}(f)\big]\subset\mathbb{R}^n$ の閉区間集合の両端点に関して，$\underline{L}(f):\mathbb{R}_+^n\to\mathbb{R}_+^n, \overline{L}(f):\mathbb{R}_+^n\to\mathbb{R}_+^n$ の写像を，$\boldsymbol{x}\in\mathbb{R}_+^n$ に対して

$$\underline{L}(f)\boldsymbol{x}=\boldsymbol{r}(f)+\beta\min_{Q\in\mathcal{Q}(f)}Q\boldsymbol{x} \quad (4)$$

$$\overline{L}(f)\boldsymbol{x}=\boldsymbol{r}(f)+\beta\max_{Q\in\mathcal{Q}(f)}Q\boldsymbol{x} \quad (5)$$

とおくとき，$\underline{L}(f)$ と $\overline{L}(f)$ に関する $\mathbb{R}_+$ 上の不動点としてそれぞれ $\underline{\phi}(f),\overline{\phi}(f)$ が得られる．最適政策 f^* は，より高い上限値が得られる決定 $a(i)$ を定常政策とするものを考えれば良く，上限値に関する式 (5) の作用素 $\overline{L}$ での Maximizer $f^*(i)$: $f^*(i)=\arg\max_{a\in A}\big\{\overline{r}(i,a)+\beta\max_q\sum_j q(j|i,a)\overline{x}_j\big\}$ によって得られる．ただし，$q=q(j|i,a)$ は，状態 $i\in S$，決定 $a\in A$ に対する区間型推移確率行列 $Q\in\mathcal{Q}$ の第 i 行目である．さらに，たとえば，利得関数とコスト関数の各成分 $r(i,a), c(i,a)$ が p 次元であるとき，縮小写像とその不動点定理によって，閉凸錐 $K\subset\mathbb{R}^p$ の半順序によるパレート最適性での最適解（最適政策）$f^{*(\infty)}=(f^*,f^*,\ldots)\in\Pi$ が存在することが示される（これらのことは，詳しくは，Kurano et al. [9] を参照）．

3. 推移確率行列の区間ベイズ推定

各状態 $i\in S$ での決定 $f_t(i)\in A$ による推移確率ベクトル $q(i,f(i))=\big(q(1|i,f_t(i)),\ldots,q(n|i,f_t(i))\big)$ を第 i 行目にもつ推移確率行列 $Q(f_t)$ が，各時刻 t での状態 X_t の推移法則である．したがって，確率ベクトルの推定にどのような手法を用いるかが不確実性下での MDPs 構成のためのはじめの課題となる．また，区間ベイズ手法による推測区間 $\big[\underline{Q}_t,\overline{Q}_t\big]$ は時刻 t に関して真の推移確率行列 Q に漸近的に収束することが示される [10] から，ここでは，定常マルコフ決定過程，すなわち定常政策での区間型 MDPs を考える．推定する $Q(f)$ の第 i 行目の確率ベクトル $q(i,f(i))$ の全体を $P_n:=P(S)=\big\{p=(p_1,\ldots,p_n)|p_i\geqq 0,\sum_{i\in S}p_i=1\big\}$ とおく．$\mathcal{B}$ を $\mathbb{R}^n$ 上の可測集合の全体とし，$\mathcal{B}$ 上の測度 L,U に対し $L\leqq U$ であるとは，任意の $A\in\mathcal{B}$ に対し $L(A)\leqq U(A)$ を満たす場合とする．また，測度に関する凸集合 $[L,U]$ とは，測度 Q で任意の $A\in\mathcal{B}$ に対して，$L(A)\leqq Q(A)\leqq U(A)$ を満たすものの全体である．$[L,U]$ を事前測度区間とし，ベイズ手法により事後測度区間から区間型推移確率行列を構成する MDPs を区間推定マルコフ決定過程と呼ぶ．議論を簡単にするために，ルベーグ測度 L と定数 $k\geqq 1$ を用いて $[L,kL]$ を事前測度区間とおく．データセット $\sigma=(\sigma_1,\ldots,\sigma_n)$ は，総試行回数 N の独立試行のもとで，

各状態 j への推移回数の観測値が σ_j である成分からなる．$p=(p_1,\ldots,p_n)\in P_n$ の各パラメータ p_i の推定を観測データ $\sigma=(\sigma_1,\ldots,\sigma_n)$ によって行うとき，次の多項分布の確率関数 f:

$$f(\sigma_1,\ldots,\sigma_n|p)=\frac{(\sigma_1+\cdots+\sigma_n)!}{\sigma_1!\cdots\sigma_n!}p_1^{\sigma_1}\cdots p_n^{\sigma_n} \quad (6)$$

で推定を行う．このとき，$[L,U]$ と σ による事後測度区間を $[L_\sigma,U_\sigma]$ と表せば，p_i の事後測度区間 $[\underline{\lambda}_i,\overline{\lambda}_i]$ は，次の積分比による事後期待測度区間として表される：$\left\{\int_{P_n}p_i\,Q(dp)/\int_{P_n}Q(dp)\middle|L_\sigma\leqq Q\leqq U_\sigma\right\}$．また，下限値 $\underline{\lambda}_i$ と上限値 $\overline{\lambda}_i$ は，それぞれ次の積分方程式の一意の解である：$U_\sigma(p_i-\underline{\lambda}_i)^-+L_\sigma(p_i-\underline{\lambda}_i)^+=0, U_\sigma(p_i-\overline{\lambda}_i)^++L_\sigma(p_i-\overline{\lambda}_i)^-=0$. ただし，実数 x に対して，$x^+=\max\{0,x\}, x^-=x-x^+=\min\{0,x\}$ である．多項分布のパラメータ p に関する確率分布はディリクレ分布であり，上記二つの積分方程式の解法は，事前測度区間 $[L,kL]$ のとき，次の各方程式の不動点探索（方程式の零点探索）に等しい：

$$\underline{\lambda}_i=\frac{B(s+1,t)+(k-1)B(s+1,t,\underline{\lambda}_i)}{B(s,t)+(k-1)B(s,t,\underline{\lambda}_i)} \quad (7)$$

$$\overline{\lambda}_i=\frac{kB(s+1,t)-(k-1)B(s+1,t,\overline{\lambda}_i)}{kB(s,t)-(k-1)B(s,t,\overline{\lambda}_i)} \quad (8)$$

ただし，$s=\sigma_i+1, t=\sum_{j=1}^n\sigma_j-\sigma_i+(n-1)=N-\sigma_i+(n-1)$, $B(s,t)=\int_0^1x^{s-1}(1-x)^{t-1}dx$, $B(s,t,\lambda)=\int_0^\lambda x^{s-1}(1-x)^{t-1}dx$ である．

4. OERA モデルへの適用例

ここでは，Cover and Keilers [11] の提案する選手の打撃力評価について区間推定マルコフ決定過程を適用する．状態空間を $S=\{0,1,2,\ldots,24\}$ とし，状態 0 は 3 アウトの状態を表す．状態 1 から 8 はノーアウトであり，状態 1 はランナーなし，状態 2 はランナー 1 塁，状態 3 はランナー 2 塁，状態 4 はランナー 3 塁，状態 5 はランナー 1, 2 塁，状態 6 はランナー 1, 3 塁，状態 7 はランナー 2, 3 塁，状態 8 は満塁，以下，各状態番号についてアウトカウントの増加に応じて考える．Howard [12] (Chap.5) のように表現すると，表 1 のように表せる．たとえば，状態 2, 10, 18 はそれぞれのアウトカウントが 0, 1, 2 であって，1 塁にランナーがいて (1), 2, 3 塁上にはランナーがいない（それぞれ 0 である）ことを示す．

OERA モデルでは，次のようなルールのもとで確率過程の状態推移が進行する：(i) 選手ごとに，推移確率行列 P (25×25 行列) がある．(ii) $P=(p_{ij})$ について，各成分 p_{ij} は，0 の値または 6 個のパラメータ

表 1　状態番号と塁上のランナーの配置状況

State	Outs	Bases 1 2 3	State	Outs	Bases 1 2 3
0	3	- - -			
1,9,17	0,1,2	0 0 0	5,13,21	0,1,2	1 1 0
2,10,18	0,1,2	1 0 0	6,14,22	0,1,2	1 0 1
3,11,19	0,1,2	0 1 0	7,15,23	0,1,2	0 1 1
4,12,20	0,1,2	0 0 1	8,16,24	0,1,2	1 1 1

値 p_O（凡打），p_B（四死球），p_1（1 塁打），p_2（2 塁打），p_3（3 塁打），p_4（本塁打）のいずれかによって定まる．(iii) 打席に立つ選手の推移法則 P の打席結果でランナーの進塁が生じる．進塁には，次のような所定のルールがある．1 塁打と 2 塁打は長打とし，1 塁打では塁上のランナーは二つ進塁する．2 塁打では 1 塁上のランナーも生還し得点となる．また，この選手評価モデルでは犠打を扱わず，野手のエラーはアウトカウント，ダブルプレーはないものとしている．

推移確率行列 P は，次のような小行列で表され，P による状態推移が吸収マルコフ連鎖をなす確率過程の数理モデルがつくられる．すなわち，$P=\begin{pmatrix}1 & \mathbf{0}\\ \boldsymbol{R}' & Q\end{pmatrix}$ であって，$\mathbf{0}$ と $\boldsymbol{R}$ は，それぞれ 1×24 のすべての成分が 0 である零行列（行ベクトル）と非負値行列（行ベクトル）であり，$\boldsymbol{R}'$ は $\boldsymbol{R}$ の転置行列（列ベクトル）である．上記の過渡的状態の集合 $S_1\cup S_2\cup S_3$ に対応する小行列 Q（24×24 行列）に関しては，単位行列 I との差である $I-Q$ が正則であるとき，次の等式を満たす基本行列 $(I-Q)^{-1}$ が知られている．

$$\begin{aligned}&W:=I+Q+Q^2+Q^3+\cdots,\\ &W=I+QW\\ \therefore\quad &W=(I-Q)^{-1}\end{aligned}$$

さらに，吸収状態 0 への到達時刻（停止時刻）を表す確率変数を τ とすると，

$$W=(w_{ij}), w_{ij}=E\left[\textstyle\sum_{n=0}^{\tau-1}I_{\{X_n=j\}}|X_0=i\right]$$

であって，状態 $i\in S-\{0\}$ から出発して状態 0 に吸収されるまでの平均吸収時間 $m_i=E[\tau|X_0=i]$ について Pinsky and Karlin [13] (p. 142) より $\sum_{j=1}^{24}\sum_{n=0}^{\tau-1}I_{\{X_n=j\}}=\sum_{n=0}^{\tau-1}\sum_{j=1}^{24}I_{\{X_n=j\}}=\sum_{n=0}^{\tau-1}1=\tau$ が得られるから，$\sum_{j=1}^{24}w_{ij}=E[\tau|X_0=i]=m_i$ $(i\in S-\{0\})$ となり，基本確率行列 $(I-Q)^{-1}$ のそれぞれの第 i 行目の行和 $\sum_{j=1}^{24}w_{ij}$ が平均吸収時間 m_i を表す．

一般に，停止時刻 $\{\tau=n\}\in\mathcal{F}_n=\sigma(X_0,\ldots,$

$X_n), n = 0, 1, \ldots$ に対して，最良選択問題（秘書問題）のように停止政策 τ によるマルコフ連鎖が構成されるとき，一つの τ による確率過程とそこでの価値関数（効用）の評価値が定まると考えることもできる．OERA モデルでは，ひとりの選手が 1 イニング 3 アウト，さらに 9 イニング分打席に立ち続けるモデルで，そこへ停止政策として塁上のランナー配置状況の場面ごとに代打を考慮し状態推移を 3 アウトの前であっても停止できる打撃力評価もマルコフ決定モデル（停止決定モデル）として解析可能である．私にとって，学部生への研究指導の“テッパンネタ”の一つが，本稿のマルコフモデルの分析である．セイバーメトリクス（たとえば，鳥越ら [14]）のように多面的に幅広く分析を行う方法もある．オペレーションズ・リサーチ界隈での野球分析の関連話題は，穴太克則先生 [15]，吉良知文先生 [16]，廣津信義先生 [17] のマルコフモデルによる研究もそれぞれある．興味ある読者諸氏には，これらの文献もお薦めする．脱線ついでに，野球に関する話題をはじめ，待ち行列，スケジューリング，レベニューマネジメント，ランキング，ヘルスケアの話題など，私の周りの学生が興味をもって勉強している好適書の高木英明編著 [18, 19] も紹介する．

5. 具体例（OERA とヘルスケア）

OERA モデルでの分析例では，状態空間を $S = \{0, 1, \ldots, 24\} = \{0\} \cup \{1, \ldots 8\} \cup \{9, \ldots 16\} \cup \{17, \ldots 24\} := \{0\} \cup S_1 \cup S_2 \cup S_3$ とおく．パラメータ $p_i, i \in \{O, B, 1, 2, 3, 4\}$ に対して，真の推移確率行列 P における Q の要素 $q_{ij}, i, j \in S_1 \cup S_2 \cup S_3$ は，次のような小行列の推移行列の要素である（cf. 木下 [20]）．

$$Q = \begin{pmatrix} Q_{11} & Q_{12} & Q_{13} \\ Q_{21} & Q_{22} & Q_{23} \\ Q_{31} & Q_{32} & Q_{33} \end{pmatrix}$$

に対して，各 $Q_{ij}, i, j \in \{1, 2, 3\}$ は 8×8 の非負行列であり，具体的には次のような成分をもつ．

$$Q_{11} = \begin{pmatrix} p_4 & p_1 + p_B & p_2 & p_3 & 0 & 0 & 0 & 0 \\ p_4 & 0 & p_2 & p_3 & p_B & p_1 & 0 & 0 \\ p_4 & p_1 & p_2 & p_3 & p_B & 0 & 0 & 0 \\ p_4 & p_1 & p_2 & p_3 & 0 & p_B & 0 & 0 \\ p_4 & p_0 & p_2 & p_3 & 0 & p_1 & 0 & p_B \\ p_4 & p_0 & p_2 & p_3 & 0 & p_1 & 0 & p_B \\ p_4 & p_1 & p_2 & p_3 & 0 & 0 & 0 & p_B \\ p_4 & 0 & p_2 & p_3 & 0 & p_1 & 0 & p_B \end{pmatrix}$$

先述の進塁方法のルールから $Q_{11} = Q_{22} = Q_{33}$ を満たし，Q_{12} と Q_{23} は，対角成分がすべて p_O の対角行列，残りの $Q_{ij}, i < j$ については $Q_{21} = Q_{31} = Q_{32} = O = (0)$（零行列）である．また，$P = \begin{pmatrix} 1 & \mathbf{0} \\ \boldsymbol{R}' & Q \end{pmatrix}$ に対して，$\boldsymbol{R} = (0, \ldots, 0, p_O, p_O, p_O, p_O, p_O, p_O, p_O, p_O)$ で最後の 8 個の各成分が p_O の行ベクトルである．データセット σ による推定された区間表現は，$p_i \in [\underline{p}_i, \overline{p}_i], i \in \{O, B, 1, 2, 3, 4\}$ であり，それらにもとづいて区間型推移確率行列 $\langle L, U \rangle = \left\{ P = (p_{ij}) \in \mathbb{R}_+^{25 \times 25} \middle| \ \underline{l}_{ij} \leqq p_{ij} \leqq \overline{u}_{ij}, p_{ij} \geqq 0, \sum_{j \in S} p_{ij} = 1 \ (i, j \in S) \right\}$ による MDPs を構成する．利得（得点）ベクトル $\boldsymbol{r}$ を 25×1 行列として列ベクトルで表すとき，転置ベクトル（行ベクトル）で，$\boldsymbol{r}' = (0, \boldsymbol{r_1}, \boldsymbol{r_1}, \boldsymbol{r_1})'$, $\boldsymbol{r_1}' = (p_4, 2p_4 + p_2 + p_3, 2p_4 + p_1 + p_2 + p_3, 2p_4 + p_1 + p_2 + p_3, 3p_4 + 2(p_2 + p_3) + p_1, 3p_4 + 2(p_2 + p_3) + p_1, 3p_4 + 2(p_2 + p_3 + p_1), 4p_4 + 3(p_2 + p_3) + 2p_1 + p_B)'$ である．この利得ベクトル $\boldsymbol{r}$ についても事後期待測度による区間表現 $\boldsymbol{r} = [\underline{r}, \overline{r}]$ を得ることができる．一般に，Cover and Keiler [11] の OERA モデルなど完全情報マルコフ過程下では，利得計算（価値関数の評価値）を先述の基本行列を用いて，$(I - Q)^{-1}\boldsymbol{r}$ で行う．区間型マルコフモデルでは，基本行列の計算としての逆行列の計算が困難であるから，$\left(I + Q + Q^2 + \cdots + Q^k\right)\boldsymbol{r}$ による k について逐次近似計算を行うことになる．また，このとき，$\langle L, U \rangle$ は凸多面体であるから，区間型行列のべき乗の反復計算では，超平面の交点間の端点を利用した計算によって新たな凸多面体が逐次計算される．

数値例に，まず，2001 年度の読売ジャイアンツ松井秀喜選手の区間型マルコフモデルを示す（成績データは，木下 [20] より）．事前測度区間を $[L, 2L](k = 2)$ とし，データセット $\sigma = (\sigma_O, \sigma_B, \sigma_1, \sigma_2, \sigma_3, \sigma_4) = (321, 123, 98, 23, 3, 36)$，$N = 604$ であり，$p_i, i \in \{O, B, 1, 2, 3, 4\}$ の区間表現として，$p_O \in [0.5223, 0.5334]$, $p_B \in [0.1988, 0.2078]$, $p_1 \in [0.1582, 0.1664]$, $p_2 \in [0.0371, 0.0415]$, $p_3 \in [0.0057, 0.0075]$, $p_4 \in [0.0580, 0.0633]$ を得る．また，$p_B + p_1$ の成分推定を $\sigma_B + \sigma_1 = 221$ で行うと，$p_B + p_1 \in [0.3586, 0.3693]$ を得る．たとえば，状態 1（ノーアウトランナーなし）での平均吸収時間 m_1 は，$k = 21$ として $I + Q + \cdots + Q^k$ の第 1 行目の行和から $m_1 \in [4.4816, 4.8802]$ と見積もられる．状態 1 から出発し状態 0 に吸収されるまでの各状態 $j (0 < j < 25)$ への平均訪問回数 m_{1j} を図 1 に示す．

さらに，Baseball savant[21] から，松井秀喜選手

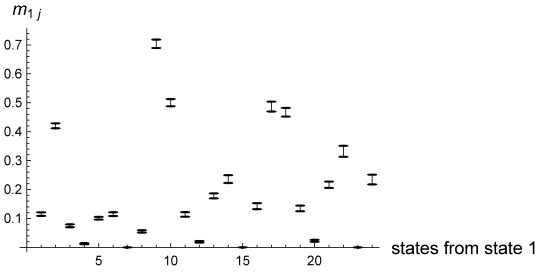

図 1　状態 1 からの平均訪問回数 $m_{1j}(0 < j < 25)$

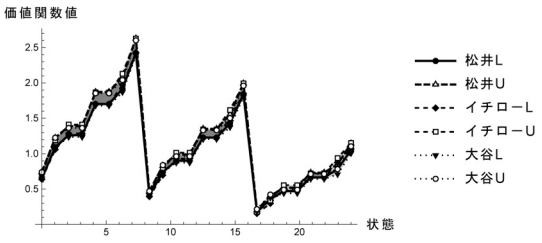

図 2　3 選手の価値関数値の区間比較

表 2　$A = \{1$（松井），2（イチロー），3（大谷）$\}$ での期待得点と最適政策での期待得点 $(\beta = 0.9)$（ノーアウトの各状態の場合）

状態	$f^*(i)$ と $\phi(f^*)$
1	1, 1, $[0.6517, 0.7410]$
2	3, 3, $[1.0940, 1.2279]$
3	2, 2, $[1.2894, 1.4098]$
4	2, 2, $[1.2894, 1.4098]$
5	1, 1, $[1.7012, 1.8720]$
6	1, 1, $[1.7012, 1.8720]$
7	2, 2, $[1.9726, 2.1301]$
8	2, 1, $[2.4332, 2.6301]$

（ニューヨークヤンキース，2004 年，レギュラー + ポストシーズン），イチロー選手（シアトルマリナーズ，2004 年，レギュラーシーズン），大谷翔平選手（ロサンゼルスエンジェルス，2021 年，レギュラーシーズン）の各打撃成績から区間推定マルコフ決定過程の各パラメータ値を区間推定し，過渡的状態 $i \in S$ の割引評価規準での期待得点の上限値，下限値の概形を図 2 に示す．さらに，状態 $i \in S$ でのノーアウトに対する最適戦略 $f^*(i) = (\underline{f}^*(i), \overline{f}^*(i))$ と価値関数の区間値 $\phi(f^*) = [\underline{\phi}(f^*), \overline{\phi}(f^*)]$（1 イニング）をまとめたものを表 2 に示す．

ベクトル集合値関数での最適化に関しては，Furukawa [22] に政策改良法によるアプローチ研究がある．また，野球と MDPs の研究に関しては，先述の Howard [12] の研究をはじめとして 数多くあるが，分析モデルに物足りなさを感じる読者諸氏も少なくないと思われる．たとえば，オペレーションズ・リサーチ (Vol.24, No.6) に Ladaney and Machol 編著の Optimal Strategies in Sports(1977) に関する書評 [23] があり，いつの時代にも OR によるスポーツ分析のアイデアが豊富にあることを物語っている．本稿でのメジャーリーグでの 3 選手の起用についての決定空間の変数選択による真の推移確率行列は 3^{24} 通りあるから，区間推定モデルでの最適化においても，前述の中出 [4] や大野 [2, 3] の書籍にも挙げられている各種アルゴリズムや近似手法 (Approximate Dynamic Programming) の適用や改良が必要である．

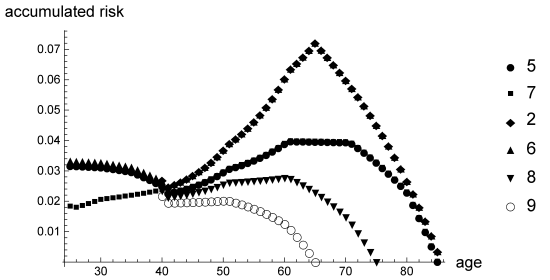

図 3　各シナリオでの検診開始時点からの累積死亡リスク

ヘルスケアへのベイズ手法によるマルコフ決定過程での応用分析事例としては，乳がん検診のシナリオ評価を挙げておく．国内の検診と KapWeb [24] での生存確率の諸データにもとづいて，部分観測マルコフ決定モデル (POMDPs) を構成し，検診結果と罹患状態推定にベイズ手法を用いた分析を行った [25]．国内の指針では，40 歳以上で 2 年に 1 回の乳がん検診の受信が推奨されており，ここでの研究で扱ったシナリオによるマルコフモデルによる評価では，1 年ごとの推移確率行列として，区分的（5 年ごと）に推移確率行列が変化する非定常マルコフ連鎖のもと，検診群と非検診群の死亡リスクの比較を行った．想定した比較対象のシナリオのもとで，推奨年齢期間での定期受診が一番死亡リスクを下げられるという結果を得た．図 3 において，シナリオ番号ごとに，5（25–39 受診なし，40–84 受診あり），7（25–84 受診あり），2（40–64 受診あり，65–84 受診なし），6（25–39, 65–84 受診なし，40–64 受診あり），8（40–84 受診あり），9（40–64 受診あり）を表す．受診の継続は，死亡リスクを増大させない，若年層 (25–39) での受診の有無は 40 歳以降の受診群の死亡リスクを増大させない，未受診が死亡リスクを高めることがこのマルコフモデルでも確認できる．

謝辞　最後に，本稿を提出する機会を下さりました大阪公立大学北條仁志先生に深謝申し上げます．

参考文献

[1] 後藤昌司，数学講座 マルコフ型逐次決定過程 (1),(2),(3)，オペレーションズ・リサーチ：経営の科学，**13**，(2) pp. 53–57，(3) pp. 40–44，(5) pp. 59–64，1969.
[2] 大野勝久，『Excel による生産管理』，朝倉書店，2011.
[3] 大野勝久，『サプライチェーンの最適運用』，朝倉書店，2011.
[4] 中出康一，『マルコフ決定過程—理論とアルゴリズム—』，コロナ社，2019.
[5] R. J. Boucherie and N. M. Dijk (eds.), *Markov Decision Processes in Practice*, Springer, 2017.
[6] O. L. V. Cosra and F. Dufour, *Continuous Average Control of Piecewise Deterministic Markov Processes*, Springer, 2013.
[7] A. Piunovskiy and Y. Zhang (eds.), *Continuous-Time Markov Decision Processes*, Springer, 2020.
[8] D. J. Hartfiel, *Markov Set-Chains: Volume 1695 of Lecture Notes in Mathematics*, Springer-Verlag, 1998.
[9] M. Kurano, J. Nakagami and M. Horiguchi, "Controlled Markov set-chains with set-valued rewards," In *Proceedings of the International Conference on Nonlinear Analysis and Convex Analysis (NACA98)*, W. Takahashi and T. Tanaka (eds.), World Scientific, pp. 205–212, 1999.
[10] L. De Robertis and J. A. Hartigan, "Bayesian inference using intervals of measures," *The Annals of Statistics*, **9**, pp. 235–244, 1981.
[11] T. M. Cover and C. W. Keilers, "An offensive earned-run average for baseball," *Operations Research*, **25**, pp. 729–740, 1977.
[12] R. A. Howard, *Dynamic Programming and Markov Processes*, The Technology Press of M.I.T., 1960.
[13] M. A. Pinsky and S. Karlin, *An Introduction to Stochastic Modeling, 4th ed.*, Academic Press, 2011.
[14] 鳥越規央，『データスタジアム事業部，勝てる野球の統計学』，岩波書店，2014.
[15] 武井貴裕，瀬古進，穴太克則，"野球の最適打順を考えてみよう，"オペレーションズ・リサーチ：経営の科学，**47**，pp. 142–147，2002.
[16] 吉良知文，稲川敬介，"野球への動的計画アプローチ，"オペレーションズ・リサーチ：経営の科学，**59**，pp. 378–384，2014.
[17] N. Hirotsu and M. Wright, "A Markov chain approach to optimal pinch hitting strategies in a designated hitter rule baseball game," *JORSJ*, **46**, pp. 353–371, 2003.
[18] 高木英明編著，『サービスサイエンスことはじめ』，筑波大学出版会，2014.
[19] 高木英明編著，『サービスサイエンスの事訳』，筑波大学出版会，2017.
[20] 木下栄蔵，『Q&A で学ぶ確率・統計の基礎』，講談社，2003.
[21] Baseball Savant, https://baseballsavant.mlb.com（2022 年 1 月 7 日閲覧）
[22] N. Furukawa, "Characterization of optimal policies in vector-valued Markovian decision processes," *Mathematics of Operations Research*, **5**, pp. 271–279, 1980.
[23] 増田伸爾，"書評：Optimal Strategies in Sports（スポーツの最適戦略），"オペレーションズ・リサーチ：経営の科学，**24**，p. 380，1979.
[24] KapWeb, Survival statistics of Japanese association of Clinical Cancer Centers, https://kapweb.chiba-cancer-registry.org（2021 年 2 月 21 日閲覧）
[25] M. Horiguchi, "On an approach to evaluation of health care programme by Markov decision model," *Modern Trends in Controlled Stochastic Processes: Theory and Applications, V.III*, A. Piunovskiy, Y. Zhang (eds.), Springer, pp. 341–354, 2021.
[26] S. S. Wilks, *Mathematical Statistics*, John Wiley & Sons, 1962.（田中英之，岩本誠一訳，『数理統計学・増訂新版 1, 2』，東京図書，1971, 1972.）

