数理解析研究所講究録
第726巻 1990年 13-26

複数のコスト制約を持つsemi-Markov決定過程

京都大学 工学部 大西 匡光 (Masamitsu Ohnishi)

§1 はじめに

よく知られているように semi-Markov決定過程（Semi-Markov Decision Process：ＳＭＤＰ，Markov決定過程（Markov Decision Process：ＭＤＰ）を含む）は待ち行列システムの制御，在庫管理，信頼性システムの保全など，確率的システムの動的最適化を考える際，有効な数学的手法であり，その最適政策を求める計算アルゴリズムに関する研究も進んでおり，政策反復法（政策改良法），値反復法（逐次近似法），線形計画問題として定式化し Simplex法を用いる解法など，かなりに大規模な問題を扱うにも耐え得るアルゴリズムも開発されて来ている．

一方 通常のＳＭＤＰではシステムの評価尺度としてただ１つのみを許しており，様々の評価尺度を考慮する必要のある現実問題に適用する際に困難さがある．例えば複数のクラスの客を有する待ち行列システム（客のクラスが異なればシステムへの要求も異なる）における動的サービス割り当て問題をＳＭＤＰで定式化するにはそれぞれのクラスの客を待たせることによる不都合を１つの評価尺度により相対的にせよ数量化（例えばコスト評価）する必要がある．また在庫管理問題では品物の発注コスト，在庫の保管に関しては倉庫の維持に要するコストおよび品物の腐敗あるいは品質の劣化による損失などを含む保管コスト，さらには品切れによって客の信用を失うことによる不利益などを表す品切れ損失を１つの評価尺度（例えばコスト）で統一的に評価し，それらの和を最小化するという形の問題にする必要がある．同様に信頼性システムの保全問題ではアベイラビリティ，保全コストなど，種類の異なる評価尺度が存在するが，例えばシステムの故障による不利益をダウンペナルティの形でコスト評価し，それと保全に要するコストとの和を最小化することを目的とすることがほとんどである．

このように複数の評価尺度を要する確率的システムの動的最適化を目標として

（１）それぞれの評価尺度を目的関数とする多目的ＳＭＤＰ，

（２）１つの評価規範を目的関数とし，他の評価規範を制約の形で取り入れた制約を持つＳＭＤＰ，

の研究が最近になり活発になされてきている．

本報告では特に（２）の制約を持つＳＭＤＰに関して最近著者が得た理論的結果，計算アルゴリズムを中心に述べ，さらに応用に関する話題にも触れる．

§2 複数種類のコストを持つSMDP

[複数種類のコストを持つ有限SMDP]

$S := \{1,2,\ldots,M\}$：状態空間，

$A(i)$：状態i（$\in S$）でとり得るアクションの有限集合，

$A := \bigcup_{i\in S} A(i)$：アクション空間，

$c_n(i,a)$ $(n=0,1,\ldots,N)$：状態i（$\in S$）でアクションa（$\in A(i)$）をとった時，次の状態遷移までに課せられる第n種の期待即時コスト，

$d(i,a)$：状態i（$\in S$）でアクションa（$\in A(i)$）をとった時，次の状態遷移までの期待時間間隔（$0 < d(i,a) < \infty$（$i\in S, a\in A(i)$）を仮定する），

$p(i,a,j)$：状態i（$\in S$）でアクションa（$\in A(i)$）をとった時，次に状態j（$\in S$）に遷移する確率．□

[履歴と政策]

$h_t := (x_0,a_0,d_1,x_1,a_1,d_2,x_2,\ldots,x_{t-1},a_{t-1},d_t,x_t)$：第$t$（$=0,1,2,\ldots$）回目の意思決定時における過去の履歴，ここで x_t,a_t,d_{t+1}（$t=0,1,2,\ldots$）は第t回目の状態，アクション，および遷移時間間隔，

r_t：第t（$=0,1,2,\ldots$）回目の意思決定時における意思決定規則，すなわち，履歴h_tの時にアクションa（$\in A(x_t)$）を選択する確率$r_t(\{a\}|h_t)$を規定する条件付き確率測度，

$R := (r_0,r_1,\ldots,r_t,\ldots)$：政策．□

[政策のクラス]

C　　：すべての政策からなる集合，

C_M（$\subset C$）：Markov政策，すなわち，各r_tがtおよびx_tを通してのみ(h_t,x_t)に依存する政策Rからなる集合，

C_S（$\subset C_M$）：定常政策，すなわち，各r_tがx_tを通してのみh_tに依存する政策Rからなる集合，

C_D（$\subset C_S$）：純粋政策，すなわち，定常であり，かつ確定的な政策からなる集合（C_Dは $\prod_{i\in S} A(i)$と同一視できることから純粋政策Rは

$$R = (R(1),R(2),\ldots,R(M)) \qquad (2.1)$$

と書ける，ただし$R(i)$（$\in A(i)$）は政策Rが状態i（$\in S$）でとるアクションを表す）．□

以下では次の仮定のもとで議論を行う．

[仮定]《単一連鎖（Unichain）の仮定》

任意の純粋政策R（$\in C_D$）に対し，Rにより導かれる時間斉次なsemi-Markov過程（あるいは時間斉次なMarkov Renewal過程）に埋め込まれた時間斉次なMarkov連鎖（その1ステップ遷移確率行列は$[p(i,R(i),j);i,j\in S]$となる）はただ1つの再帰類を持つ．□

この仮定は「任意の純粋政策R（$\in C_D$）に対して，Rにより導かれる時間斉次なMarkov連鎖においては他のすべての状態から到達可能な（Rに依存した）状態が存在する．」と言い替えることもできる．

§3　コスト制約の無い平均コスト規範ＳＭＤＰ

[Ｎ＋１種の平均コスト]

初期状態をi（$\in S$），政策をR（$\in C$）とした時の第n（$=0,1,\ldots,N$）種のコストに関する無限計画期間における長時間平均の単位時間当りの期待コスト（以下では単に平均コストと言う）$g_n(i,R)$はT（$=0,1,2,\ldots$）に対し

$$g_n^T(i,R) := \frac{E_R[\sum_{t=0}^{T} c_n(X_t,A_t)|X_0=i]}{E_R[\sum_{t=0}^{T} d(X_t,A_t)|X_0=i]} = \frac{\sum_{t=0}^{T}\sum_{j\in S}\sum_{a\in A(j)} P_R[X_t=j,A_t=a|X_0=i]c_n(j,a)}{\sum_{t=0}^{T}\sum_{j\in S}\sum_{a\in A(j)} P_R[X_t=j,A_t=a|X_0=i]d(j,a)} \quad (3.1)$$

として

$$g_n(i,R) := \limsup_{T\to+\infty} g_n^T(i,R) \quad (3.2)$$

で定義される，ただし$P_R[\cdot]$および$E_R[\cdot]$は政策Rを用いた時の確率測度とその期待値を表し，X_tおよびA_tは第t（$=0,1,2,\ldots$）回目の状態とアクションを表す確率変数である．□

[注３．１]

（１）定常政策R（$\in C_S$）のもとでの平均コスト$g_n(i,R)(n=0,1,\ldots,N,i\in S)$は初期状態iに依存せず

$$g_n(R) := g_n(i,R) = \sum_{j\in S}\sum_{a\in A(j)} \nu(j,a,R)c_n(j,a) \quad (3.3)$$

で求められる，ただし$\nu(i,a,R)$（$i\in S$，$a\in A(i)$）は政策Rのもとで状態iを訪れさらにアクションaをとる時刻の長時間平均の単位時間当りの期待頻度，あるいはその平均再帰時間の逆数である．

定常政策R（$\in C_S$）に対して平均コストベクトルを

$$g(R) := (g_0(R),g_1(R),\ldots,g_N(R))$$

と定義する.

（２）純粋政策R（$\in C_D$）に対し

$$\nu(i,R) := \nu(i,R(i),R) \quad \text{for } i\in S$$

$$\nu(R) := (\nu(i,R);i\in S)$$

と定義すると，$\nu(i,R)$は政策Rにより導かれる時間斉次なsemi-Markov過程における状態iの平均再帰時間の逆数である.

$\nu(R)=(\nu(i,R);i\in S)$の計算は以下のいずれの方法でも実行され得る.

（２－１）連立１次方程式

$$\sum_{i\in S} \nu(i,R)p(i,R(i),j) = \nu(j,R) \quad \text{for } j\in S \tag{3.4}$$

$$\sum_{i\in S} \nu(i,R)d(i,R(i)) = 1 \tag{3.5}$$

を解く.

（２－２）連立１次方程式

$$\sum_{i\in S} \pi(i,R)p(i,R(i),j) = \pi(j,R) \quad \text{for } j\in S \tag{3.6}$$

$$\sum_{i\in S} \pi(i,R) = 1 \tag{3.7}$$

を解き，政策Rにより導かれる時間斉次なsemi-Markov過程に埋め込まれた時間斉次なMarkov連鎖の定常分布$\pi(R)=(\pi(i,R);i\in S)$を求め，

$$\nu(i,R) = \frac{\pi(i,R)}{\sum_{j\in S} \pi(j,R)d(j,R(j))} \quad \text{for } i\in S \tag{3.8}$$

で得る. □

[ＳＭＤＰ$_0$]《コスト制約の無いＳＭＤＰ》

第０種の平均コスト$g_0(i,R)$をすべての初期状態i（$\in S$）に対し最小化する政策R（$\in C$）を求めよ. □

以下はＳＭＤＰにおいて良く知られている結果である（例えば Ross [11], Tijms [14]を見よ）.

[定理３．１]

最適平均コスト

$$\inf_{R\in C} g_0(i,R) \tag{3.9}$$

は初期状態i（$\in S$）に依存せず一定である（それをg_0とおく）．さらに状態空間S上のある関数$(v_0(i);i\in S)$が存在して，次の最適性方程式を満たす：

$$v_0(i) = \min_{a\in A(i)} \{c_0(i,a)-g_0d(i,a)+\sum_{j\in S} p(i,a,j)v_0(j)\} \quad \text{for } i\in S \tag{3.10}$$

各状態i（$\in S$）に対し，式(3.10)の右辺のminを達成する（任意の）アクションをとる（任意の）純粋政策R^*（$\in C_D$）は最適である．ただし$(v_0(i);i\in S)$は最適政策のもとでの十分大きな時刻までの期待総コストに初期状態の差異が与える漸近的影響を表す相対値関数と呼ばれる未知の関数である. □

最適性方程式(3.10)の解法，最適政策の計算法としては
（１）政策反復法（政策改良法）
（２）値反復法（逐次近似法）
（３）ＬＰ問題に定式化することによる解法
などが代表的である．本稿では後での対応のため（３）のＬＰ問題に定式化することによる解法について簡単に述べておく．最適性方程式(3.10)における$g_\emptyset$および$(v_\emptyset(i)\,;\, i\in S)$は決定変数をgおよび$(v(i)\,;\, i\in S)$とする次のＬＰ問題の最適解として求めることができる．

[ＬＰ$_\emptyset$]

Maximize g

subject to

$$v(i) \leqq c_\emptyset(i,a)-gd(i,a)+\sum_{j\in S} p(i,a,j)v(j) \quad \text{for } i\in S \text{ and } a\in A(i).\square \tag{3.11}$$

ＬＰ$_\emptyset$ の双対問題は決定変数を$(x(i,a)\,;\, i\in S, a\in A(i))$とする次ののＬＰ問題となる．

[ＤＬＰ$_\emptyset$]

$$\text{Minimize} \quad \sum_{i\in S}\sum_{a\in A(i)} c_\emptyset(i,a)x(i,a) \tag{3.12}$$

subject to

$$\sum_{i\in S}\sum_{a\in A(i)} x(i,a)p(i,a,j) = \sum_{a\in A(j)} x(j,a) \quad \text{for } j\in S \tag{3.13}$$

$$\sum_{i\in S}\sum_{a\in A(i)} d(i,a)x(i,a) = 1$$

$$x(i,a) \geqq 0 \quad \text{for } i\in S \text{ and } a\in A(i).\square \tag{3.14}$$

ＤＬＰ$_\emptyset$ の決定変数$x(j,a)$（$j\in S$，$a\in A(j)$）は状態jを訪れさらにアクションaをとる時刻の長時間平均の単位時間当りの頻度，すなわち

$$x^T(j,a;i,R) := \frac{E_R\left[\sum_{t=0}^{T} 1(\{X_t=j, A_t=a\}) \mid x_\emptyset=i\right]}{E_R\left[\sum_{t=0}^{T} d(X_t,A_t) \mid x_\emptyset=i\right]} = \frac{\sum_{t=0}^{T} P_R[X_t=j, A_t=a \mid x_\emptyset=i]}{\sum_{t=0}^{T}\sum_{j\in S}\sum_{a\in A(j)} P_R[X_t=j, A_t=a \mid x_\emptyset=i]d(j,a)} \tag{3.15}$$

とし

$$x(j,a;i,R) := \limsup_{T\to+\infty} x^T(j,a;i,R) \tag{3.16}$$

で定義される量に対応している．

ＤＬＰＮの最適解を$(x^*(i,a);i\in S,a\in A(i))$とすると最適政策$R^*$（$\in C_S$）は各状態i（$\in S$）において

$$b(i,a) := x^*(i,a)/\sum_{e\in A(i)} x^*(i,e) \qquad (3.17)$$

の確率でアクションa（$\in A(i)$）を選択することで実現できる．よって最適政策$R^*=(R^*(1),R^*(2),\ldots,R^*(M))$（$\in C_D$）は以下のアルゴリズムで求められる．

［Ａ１］

Step 1：

$$E \leftarrow \{i\in S:\ \sum_{a\in A(i)} x^*(i,a) > 0\}$$

とし，すべてのi（$\in E$）に対し

$R^*(i) \leftarrow$ ［$x^*(i,a) > 0$ なる任意のアクション a（$\in A(i)$）］

とする．

Step 2：

E=Sとなるまで以下を繰り返す：

ある状態j（$\in E$）とアクションa（$\in A(i)$）に対し

$p(i,a,j) > 0$

となる状態i（$\in S-E$）を定め，

$R^*(i) \leftarrow a$, $E \leftarrow E\cup\{i\}$.

E=Sならば停止；純粋政策$R^*=(R^*(1),R^*(2),\ldots,R^*(M))$（$\in C_D$）は最適である．□

§４　複数のコスト制約を持つ平均コスト規範ＳＭＤＰ

［$SMDP_N$］《Ｎ種のコストに対する制約を持つＳＭＤＰ》

第１種から第Ｎ種のコストに関する制約

$$g_n(i,R) \leqq b_n \quad \text{for } i\in S \text{ and } n=1,\ldots,N \qquad (4.1)$$

を満たす政策の中で第０種の平均コスト$g_0(i,R)$をすべての初期状態i（$\in S$）に対し最小化する政策R（$\in C$）（が存在すればそれを）を求めよ．□

［定理４．１］

$SMDP_N$の任意の実行可能な，すなわち，制約(4.1)を満たす政策R（$\in C$）に対し，やはり実行可能な定常政策R'（$\in C_S$）が存在して

$$g_0(i,R') \leqq g_0(i,R) \quad \text{for } i\in S \qquad (4.2)$$

が成り立つ．□

上の定理より考察の対象とする政策のクラスとしてC_Sに限定してもよい，従って問題$SMDP_N$は次の問題に帰着された．

[SMDP$_N$]'

第1種から第N種のコストに関する制約

$$g_n(R) \leq b_n \qquad \text{for } n=1,\ldots,N \qquad (4.1)'$$

を満たす政策の中で第0種の平均コスト$g_0(R)$を最小化する定常政策政策R ($\in C_S$) (が存在すればそれを) を求めよ. □

[定義4. 1]

$R_1, R_2, \ldots, R_k$ ($\in C_D$) を純粋政策, $(b(i,R_j); j=1,2,\ldots,k)$ ($i \in S$) を確率ベクトルとする, ただし

$$\sum_{j=1}^{k} b(i,R_j) = 1 \qquad \text{for } i \in S$$

$$b(i,R_j) \geq 0 \qquad \text{for } i \in S \text{ and } j=1,\ldots,k.$$

定常政策R ($\in C_S$) はもしRが各状態i ($\in S$) でアクションa ($\in A(i)$) を確率

$$\sum_{\{j:a=R_j(i)\}} b(i,R_j) \qquad (4.3)$$

でとるならば混合確率を$(b(i,R_j); j=1,2,\ldots,k)$ ($i \in S$) とする$R_1, R_2, \ldots, R_k$の混合政策と言う. □

[補題4. 1]

$R_1, R_2, \ldots, R_k$ (C_D) を任意の純粋政策, $\mu_1, \mu_2, \ldots, \mu_k$を

$$\sum_{j=1}^{k} \mu_j = 1$$

を満たす任意の非負の実数とする. この時, 平均コストベクトル

$$\sum_{j=1}^{k} \mu_j g(R_j)$$

は以下で定義する混合確率$(b(i,R_j); j=1,2,\ldots,k)$ ($i \in S$) を持つ$R_1, R_2, \ldots, R_k$の混合政策で実現できる:

$$b(i,R_j) := \begin{cases} \mu_j \nu(i,R_j) \Big/ \sum_{m=1}^{k} \mu_m \nu(i,R_m) & \text{if } \sum_{m=1}^{k} \mu_m \nu(i,R_m) > 0 \\ \text{任意} & \text{if } \sum_{m=1}^{k} \mu_m \nu(i,R_m) = 0. \end{cases}$$ □

[定理4. 2]

SMDP$_N$の最適政策は高々N+1個の純粋政策の混合政策で実現できる, ただしNはSMDP$_N$の制約に含まれるコストの種類の数である.

証明の概略:

Jを純粋政策の総数とし,

$$C_D = \{R_1, R_2, \ldots, R_J\}$$

とすれば, 最適政策は (理論上は) 次のLP問題の解くことで求められる.

$$\text{Minimize} \quad \sum_{j=0}^{J} \mu_j g_0(R_j)$$

$$\text{subject to} \quad \sum_{j=0}^{J} \mu_j g_n(R_j) \leqq b_n \quad \text{for } n=1,\ldots,N$$

$$\sum_{j=0}^{J} \mu_j = 1$$

$$\mu_j \geqq 0 \quad \text{for } j=0,1,\ldots,J. \ \square$$

[注4．1]

（1）N＝1に対してはBeutler and Ross [1]が一般的なアクション空間を持つＭＤＰに対して同様の結果を得ている．

（2）Beutler and Ross [2]では彼らの[1]での結果がＳＭＤＰに対しても成立することを示し，さらに2つの純粋戦略を混合する際に高々1つの状態でのみアクションの選択をランダム化すれば良いことを示した．

（3）Muro, Ohnishi and Ibaraki [7]は一般のNに対するＭＤＰに対して上の定理と同様の結果を示した．□

《ＬＰ（Linear Programming）による解法》

ＳＭＤＰ$_N$ における最適政策を計算する方法としてはＳＭＤＰ$_0$ に対するＬＰ問題ＤＬＰ$_0$ の制約にＳＭＤＰ$_N$の制約に対応する制約式を付け加えた次のＬＰ問題を解くことによる解法が考えられる．

[ＤＬＰ$_N$]

$$\text{Minimize} \quad \sum_{i\in S} \sum_{a\in A(i)} c_0(i,a)x(i,a) \tag{4.4}$$

$$\text{subject to}$$

$$\sum_{i\in S} \sum_{a\in A(i)} c_n(i,a)x(i,a) \leqq b_n \quad \text{for } n=1,2,\ldots,N \tag{4.5}$$

$$\sum_{i\in S} \sum_{a\in A(i)} x(i,a)p(i,a,j) = \sum_{a\in A(j)} x(j,a) \quad \text{for } j\in S \tag{4.6}$$

$$\sum_{i\in S} \sum_{a\in A(i)} d(i,a)x(i,a) = 1 \tag{4.7}$$

$$x(i,a) \geqq 0 \quad \text{for } i\in S \text{ and } a\in A(i). \ \square \tag{4.8}$$

ＤＬＰ$_0$の場合と同様ＤＬＰ$_N$の最適解を$(x^*(i,a); i\in S, a\in A(i))$ とすると最適政策R*（$\in C_S$）は各状態i（$\in S$）において

$$b(i,a) := x^*(i,a) / \sum_{e\in A(i)} x^*(i,e)$$

の確率でアクションa（$\in A(i)$）を選択することで実現できると予想される．しかしながらＳＭＤＰ$_N$のすべての実行可能な政策の集合とＤＬＰ$_N$の実行可能領域との対応関係は明らかではなく，上記のＬＰによる解法の正当性は必ずしも保証されない．

本報告ではLagrange乗数法を用いたアルゴリズムを提案する.

《Lagrange乗数法による解法》

[ＳＵＳＭＤＰ]

Step 0：（初期化）

実行可能な定常政策R（$\in C_S$）を見つける．もしそのような政策が存在しなければ停止；ＳＭＤＰ$_N$は不能である．さもなければRを純粋政策$R_0, R_1, \ldots, R_m$（$\in C_D$, $m \leq N$）の混合政策とした時N-m個の任意の純粋政策を付け加えて

$$D_N \leftarrow \{R_0, R_1, \ldots, R_N\}$$

とし，D_Nの各純粋政策R_jに対し

$$g_n(R_j)\ (= g_n(i, R_j) \quad \text{for all } i \in S) \qquad \text{for } n=0,1,\ldots,N$$

を計算する．k←Nとして Step 1へ.

Step 1：

次のＬＰ問題ＬＰ$_k$を解く．

$$\text{ＬＰ}_k: \quad \text{Minimize} \quad \sum_{j=0}^{k} \mu_j g_0(R_j) \tag{4.9}$$

$$\text{subject to} \quad \sum_{j=0}^{k} \mu_j g_n(R_j) \leq b_n \quad \text{for } n=1,\ldots,N \tag{4.10}$$

$$\sum_{j=0}^{k} \mu_j = 1 \tag{4.11}$$

$$\mu_j \geq 0 \qquad \text{for } j=0,1,\ldots,k. \tag{4.12}$$

$\mu^+=(\mu^+_0, \mu^+_1, \ldots, \mu^+_k)$をＬＰ$_k$の最適解とし，

$$J_k \leftarrow \{j : \mu^+_j > 0\}$$

とする．さらに

$$\lambda^* = (\lambda^*_0, \lambda^*_1, \ldots, \lambda^*_N)$$

をＬＰ$_k$の双対問題の最適解とする，ただし λ^*_0は式(4.11)，λ^*_n（n=1,...,N）は式(4.10)に対応している．

Step 2：

$$c_0(i,a) - \sum_{n=1}^{N} \lambda^*_n c_n(i,a) \quad \text{for } i \in S \text{ and } a \in A(i) \tag{4.13}$$

を（唯一の種類の）期待即時コストとして持つコスト制約の無いＳＭＤＰを解く．もしこのＳＭＤＰの最適な平均コストがλ^*_0を越えないならばStep 3へ．さもなければその最適純粋政策をR_{k+1}（$\in C_D$）とし

$$g_n(R_{k+1})\ (= g_n(i, R_{k+1}) \quad \text{for all } i \in S) \quad \text{for } n=0,1,\ldots,N$$

を計算.

$$D_{k+1} \leftarrow D_k \cup \{R_{k+1}\}$$

$$k \leftarrow k+1$$

としてStep 1へ戻る.

Step 3：

すべてのj（$\in J_k$）に対し純粋政策R_jにより導かれる時間斉次なsemi-Markov過程における各状態の平均再帰時間の逆数のベクトル

$$\nu(R_j) = (\nu(i,R_j); i \in S)$$

を求め，確率$b(i,R_j)$ $(i \in S,\ j \in J_k)$ を次のように計算する：

$$b(i,R_j) \leftarrow \begin{cases} \mu^+_j \nu(i,R_j) / \sum_{m \in J_k} \mu^+_m \nu(i,R_m) & \text{if } \sum_{m \in J_k} \mu^+_m \nu(i,R_m) > 0 \\ \text{任意} & \text{if } \sum_{m \in J_k} \mu^+_m \nu(i,R_m) = 0. \end{cases} \qquad (4.14)$$

停止；混合確率を$(b(i,R_j); j \in J_k)$ $(i \in S)$ とする$(R_j; j \in J_k)$の混合政策は$SMDP_N$ の最適政策である．□

［注４．２］

（１）アルゴリズムＳＵＳＭＤＰはＬＰにおける列生成法（column generation technique）のアイデアを用いたものである．ＬＰにおける用語を用いれば，先ずStep 0は初期実行可能基底解を求める．Step 2は第ｎ回目の反復において“列”$g(R_{N+n})$を生成する．

（２）上記の通りにアルゴリズムを実行すれば“列”の数 k（$= |D_k|$），すなわち，変数μ_jの数は各反復ごとに１つずつ増加していくことになるが，LP_kの最適解に現れない変数を捨てていくことで$|D_k|$の増加を防ぐ工夫をすることが可能である．

（３）Step 2におけるコスト制約の無いＳＭＤＰは例えば通常の政策反復法（政策改良法）で解けば良い．□

Step 0における実行可能な定常政策は，次のアルゴリズムで述べられるように，コスト制約の無いＳＭＤＰから始めて，アルゴリズムＳＵＳＭＤＰ自身を逐次用いながら，制約に取り入れるコストの種類の数を増加させていくことにより得られる．

［ＩＮＩＴＩＡＬ］

Step 0：（初期化）

$k \leftarrow 0$

$I_0 \leftarrow \{1, 2, \ldots, N\}$

R_0（$\in C_D$）：任意の純粋政策

とする．

Step 1：α_n（> 0, $n \in I_k$）を任意に定め（例えば$\alpha_n = 1/|I_k|$, $n \in I_k$とする）アルゴリズムＳＵＳＭＤＰを用いて次の$|I_0 - I_k|$種のコスト制約を持つＳＭＤＰ（P_{k+1}）の最適政策を定常政策のクラスC_Sの中で求める．その際の初期の実行可能な定常政策としてはk = 0のときはR_0，k > 0のときは前の反復における$|I_0 - I_{k-1}|$種のコスト制約を持つＳＭＤＰ（P_k）の最適定常政策R_k（$\in C_S$）を用いれば良い．

$$P_{k+1}: \quad \text{Minimize} \quad \sum_{n \in I_k} \alpha_n g_n(R)$$

$$\text{subject to} \quad g_n(R) \leqq b_n \qquad \text{for } n \in I_0 - I_k$$

P_{k+1}の最適定常政策をR_{k+1}（$\in C_S$）とし，添え字集合を

$$I_{k+1} \leftarrow \{n \in I_k : g_n(R_{k+1}) > b_n\}$$

とする．（P_{k+1}のコスト制約の添え字集合I_0-I_kは初期には空集合であるが，各反復ごとに少なくとも１つは要素の数が増加していく，すなわち，$I_{k+1} \neq I_k$．）

Step 2:

（１）もし$I_{k+1} = \phi$ならば停止；R_{k+1}は問題ＳＭＤＰ$_N$の実行可能な定常政策である．

（２）もし$I_{k+1} = I_k \neq \phi$ならば停止；問題ＳＭＤＰ$_N$は実行不能である．

（３）さもなくば　k←k+1としてStep 1へ戻る．□

アルゴリズムＩＮＩＴＩＡＬの正当性に関してはStep 2における（２）のみを吟味すれば良い．

［定理４．３］

アルゴリズムＩＮＩＴＩＡＬのあるｋ回目の反復において

$$I_{k+1} = I_k \neq \phi$$

となれば，問題ＳＭＤＰ$_N$は実行可能な政策を持たない．□

アルゴリズムＳＵＳＭＤＰはそのStep 0においてアルゴリズムＩＮＩＴＩＡＬを用いれば，有限時間の計算の後に，問題ＳＭＤＰ$_N$の最適政策を求めるか，あるいはその実行不能性を決定して停止する（ただしアルゴリズムの実行の途中で現れるコスト制約の無いＳＭＤＰを政策反復法（政策改良法）のような有限時間収束性を持つアルゴリズムを適用して解くと仮定する）．

§５　応用

この節では§４で扱ったコスト制約を持つＳＭＤＰの一般的枠組みの中には厳密には入らないがその拡張された枠組みには入る応用例を２例紹介する．

［例５．１］（セミ・マルコフ的劣化システムの保全問題：瀬川，河合，茨木[13]）

０，１，．．．，Ｋ，Ｋ＋１のＫ＋２個の状態をとり得る多状態信頼性システムを考える．状態０は新品同様の状態，状態Ｋ＋１は故障状態，その他の状態１，．．．，Ｋは中間的な劣化状態である．保全を行わなければシステムの状態はあるセミ・マルコフ過程に従い次第に劣化し，ついには故障するものとする．システムの状態の遷移が観測された直後に次の２つのアクションの内の１つをとることができる：

Ｏ：保全を行わず，稼働を継続する，

Ｍ：保全を行う，

ただし　状態０（Ｋ＋１）ではアクションＯ（Ｍ）のみをとり得るものとする．状態ｉのシステムの保全を行う場合には平均μ_iの保全時間と単位時間当りR_iの

保全コストを要し，一方保全を行わない場合には単位時間当りL_iの稼働コストを要すると仮定する．問題は無限計画期間おける長時間平均の単位時間当りの期待コストがある規定の値以下であるという制約のもとでシステムの定常アベイラビリティを最大化する保全政策を求めることである．この問題は1種類のコストに対する制約を持つＳＭＤＰとして定式され，それにより定理４．２に対応する結果を示すことができ，ある妥当な条件のもとで「最適保全政策は２つのControl Limit 政策の混合政策である３領域政策，すなわち，劣化の程度が低い状態ではアクションＯ，高い状態ではアクションＭ，それらの中間の状態ではアクションＯとＭとをある確率でランダムにとる政策となる．」ことが示される．□

［**例５．２**］（複数の客のクラスを有する離散時間待ち行列システムにおける最適サービス割り当て問題：Nain and Ross [8]）

０，１，．．．，ＫのＫ＋１種類のクラスの客が１人のサーバーからサービスを受ける離散時間の待ち行列システムを考える．待ち合い室の容量は無限大である．各クラスｋ（＝０，１，．．．，Ｋ）の客の各時刻における到着人数は再生列を成し，λ_kをその平均到着人数とする．またクラスｋの客のサービス要求量は平均$1/\mu_k$（$0<\mu_k\leqq 1$）の幾何分布に従うものとする．システムの安定条件として

$$\rho := \sum_{k=0}^{K} \lambda_k/\mu_k < 1 \qquad (5.1)$$

を仮定する．問題はすべての初期状態ｎ（：Ｋ＋１次元非負整数ベクトル）に対して制約条件

$$g_1(n,R) := \limsup_{T\to+\infty} \frac{1}{T+1} \sum_{t=0}^{T} E_R[N_0(t) \mid N(0)=n] \leqq b_1 \qquad (5.2)$$

を満たしながら

$$g_0(n,R) := \limsup_{T\to+\infty} \frac{1}{T+1} \sum_{t=0}^{T} \sum_{k=1}^{K} c_k E_R[N_k(t) \mid N(0)=n] \qquad (5.3)$$

を最小化する（割り込みを許した）動的なサービス割り当て政策Ｒを求めることである，ただし

c_k（k=1,...,K）：非負定数

$N_k(t)$（k=0,1,...,K）：時刻ｔにおけるシステム内のクラスｋの客数を表す確率変数

$N(t) := (N_0(t), N_1(t), \ldots, N_K(t))$

である．一般性を失うことなく

$$c_1\mu_1 \leqq c_2\mu_2 \leqq \ldots \leqq c_K\mu_K$$

とする．制約(5.2)のない場合，いわゆる$c\mu$-Ruleに従う政策，すなわち，

K, K-1, ..., 1, 0

のクラスの順に高い（静的）優先権を与えるサービス割り当て政策（R_0と表す）が最適であることが知られている．一方　制約(5.2)のある場合，可算無限の状態空間と１種類のコストに対する制約を持つＭＤＰとして定式化され，定理４．２に対応した結果によりj = 0,1,...,Kに対し

$K, K-1, \ldots, j+1, 0, j, \ldots, 1$

のクラスの順で高い（静的）優先権をあたえるサービス割り当て政策をR_jと表すとすると最適なサービス割り当て政策はあるj（j=1,...,K）に対しR_{j-1}とR_jとの混合政策となる．」ことが示される．□

これらの応用例の他に通信システムにおける最適化を扱った論文として Nain and Ross [9], Ross and Chen [10], Saito [12]などがある．

§6 おわりに

紙面の制約上述べることができなかったが，ＭＤＰにおいてコストの分散を評価規範として考慮した問題も実際上重要であり，Kawai [5], Kawai and Katoh [6]などで議論されている．またコスト制約を持つＳＭＤＰの在庫管理問題への応用なども実際上重要であると思われる．今後の課題であろう．

[参考文献]

[1] Beutler, F. J. and Ross, K. W., "Optimal Policies for Controlled Markov Chains with a Constraint", Journal of Mathematical Analysis and Applications, Vol. 112, pp. 236-252, 1985.

[2] Beutler, F. J. and Ross, K. W., "Time-Average Optimal Constrained Semi-Markov Decision Processes", Advances in Applied Probability, Vol. 18, pp. 341-359, 1986.

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

[4] Hordijk, A. and Kallenberg, L. C., "Constrained Undiscounted Stochastic Dynamic Programming", Mathematics of Operations Research, Vol. 22, pp. 276-289, 1984.

[5] Kawai, H. "A Variance Minimization Problem for a Markov Decision Process", European Journal of Operational Research, Vol. 31, pp. 140-145, 1987.

[6] Kawai, H. and Katoh, N. "Variance Constrained Markov Decision Process", Journal of the Operations Research Society of Japan, Vol. 30, pp. 279-291, 1987.

[7] Muro, K., Ohnishi, M., and Ibaraki, T., "Markov Decision Processes with Multiple Constraints", in the Proceedings of the Seminar on Queueing Theory and Its Applications Held at Kyoto, Japan, May11-13, 1987, pp.43-58, 1987.

[8] Nain, P. and Ross, K. W., "Optimal Priority Assignment with Hard Constraint", IEEE Transactions on Automatic Control, Vol. AC-31, pp. 883-888, 1986.

[9] Nain, P. and Ross, K. W., "Optimal Multiplexing of Heterogeneous Traffic", ACM Performance Evaluation Review, Vol. 14, pp. 100-108, 1986.
[10] Ross, K. W. and Chen, B. T., "Optimal Scheduling of Interactive and Noninteractive Traffic in Telecommunication Systems", IEEE Transactions on Automatic Control, Vol. AC-33, pp. 261-267, 1988.
[11] Ross, S. M., Applied Probability Models with Optimization Applications, Holden-Day, Inc., San Francisco, 1970.
[12] Saito, H., "Optimal Control of Variable Rate Coding in Integrated Voice/Data Packet Networks", in Proceedings on Traffic Evaluation of Information Networks Held in Sizuoka, Japan, January 25-27, pp. 223-228, 1988.
[13] 瀬川, 河合, 茨木, "コストに制約を持つセミ・マルコフ劣化システムの最適保全政策", 電子情報通信学会誌, A Vol. J 70-A. No. 12 pp. 1822-1832 1987.
[14] Tijms, H. C., Stochastic Modelling and Analysis: A Computational Approach, John Wiley and Sons, Chichester, 1986.
[15] White, D. J., "Dynamic Programming and Probabilitistic Constraints", Operations Research, Vol. 22, pp. 654-664, 1974.