リスクを考慮した逐次的意思決定

恐神　貴行

確率的な環境下での逐次的意思決定は，期待累積コストの最小化を目的として，動的計画法に基づいて行動の戦略を決めるのが標準的である．ところが，大きな損失を避けるようにリスクを考慮する逐次的意思決定では，しばしば動的計画法が適用できない．動的計画法が適用できることと，首尾一貫した逐次的意思決定ができることとは密接な関係があり，リスクを考慮した逐次的意思決定においても動的計画法が適用できることが望ましい．本稿では，反復的リスク指標を用いることで動的計画法が適用できることを示し，逐次的意思決定における反復的リスク指標の意味をロバスト最適化の観点から議論する．

キーワード：マルコフ決定過程，時間整合性，反復的リスク指標，ロバスト最適化，動的計画法

1. はじめに

逐次的意思決定とは，先のことを考えて，すなわち将来にわたるコストを小さくするように，ときどきの状態に応じて適切な行動を選んでいくことである．このような逐次的意思決定のモデルの一つにマルコフ決定過程 (MDP) がある．近年 MDP に関連する技術が発展し [1]，また MDP のパラメタを決定するためのビッグデータが整備されており，今後 MDP の実問題への応用が急速に進んでいくと考えられる．マーケティングなどの従来からの応用に加えて，機械と人とのやり取りなどにおいても MDP は本質的な役割を果たすことが期待されている [2].

逐次的意思決定においては，期待累積コストを小さくするように，行動を選んでいくのが標準的である．MDP においては，将来の状態は，過去の状態や過去の行動に依存して，確率的に決まると考える．状態の遷移が確率的であるために，将来にわたる累積のコストも確率分布を持つが，特にその期待値に注目するのが標準的なアプローチである．

ところが，今後 MDP が多種多様な実問題に応用されるにあたって，期待累積コストの最小化は必ずしも好ましくないことがある．特に，期待累積コストを犠牲にしてでも，大きな損失を避けたい場面が多く現れると考えられる．このようなリスクを考慮した意思決定のためには，期待値とは異なるリスク指標を用いる必要がある．つまり，累積コストは確率分布を持つが，その確率分布を実数値に写像することによって，その良し悪しを議論することができる．標準的には，期待値を用いて実数値に写像するが，期待値とは異なるリスク指標を用いて実数値に写像することによってリスクを考慮した意思決定が可能となる．

本稿では，逐次的意思決定において，どのようなリスク指標を用いるべきかについて解説する．特に，本特集のテーマである動的計画法に注目して議論を進める．まず，期待累積コストの最小化に適用可能であった動的計画法が，他のリスク指標については必ずしも適用できないことを示す．また，動的計画法が適用できないことの，逐次的意思決定における意味についても議論する．特に，動的計画法が適用できないリスク指標を用いると，ある種の良い逐次的意思決定ができないことを示す．次に，動的計画法の適用を可能とする，反復的リスク指標という概念を紹介する．さらに，累積コストの反復的リスク指標値を最小にすることの意味について議論する．この意味は，パラメタに不確実性を持つ MDP におけるロバストな逐次的意思決定と関連づけることによって理解する．

2. 動的計画が適用できないリスク指標

本節では，逐次的意思決定の最適化において，動的計画法が適用できない場合について議論する．例として，地点 A から地点 C まで移動しようとする旅行者を考える．図 1 は主要な地点間の旅行時間を通常時（左）と混雑時（右）に分けて示している．特に，地点 B' から地点 C までの混雑時の所要時間は 0.8 の確率で 70 分，0.2 の確率で 150 分であることを示す．旅行者は地点 A を出発する時点においては，交通状況を知らないが，混雑する確率が 0.2 であり，通常である確率が 0.8 であることを知っている．地点 A を出発して，10

おそがみ　たかゆき
IBM 東京基礎研究所
〒 135–8511 東京都江東区豊洲 5–6–52

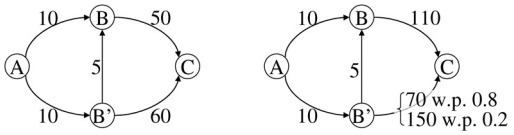

図 1　通常時（左）と混雑時（右）の旅行時間

分後に地点 B か地点 B' についたときに，交通状況を知ることができる．

この旅行者の経路選択の逐次的意思決定は MDP としてモデル化できる．状態として，（地点，交通状況）の組を考える．「地点」は旅行者のいる地点を示す．「交通状況」は，「通常」，「混雑」，「まだ交通状況を知らない」のいずれかであり，旅行者の交通状況に関する知識を示す．旅行者は各時点において，状態に応じて適切な行動（次の訪問先）を選ぶことで，与えられた目的関数を最小とするように地点 C まで到達できる．

まず標準的な期待累積コスト（時間）の最小化を考えてみる．この場合，旅行者の最適な戦略は，地点 A を出発して地点 B' を訪れ，地点 B' で交通状況を観察し，通常どおりであれば地点 B' から地点 B を経由して地点 C に到達し，混雑していれば地点 B' から地点 C に直接到達するものである．このとき期待累積時間は 71.2 分となる．

この期待累積時間を最小とする最適戦略は，動的計画法によって算出可能である．例えば，混雑時に地点 B' を出発して地点 C に向かう旅行者を考えると，期待累積時間を最小とするのは地点 B' から地点 C に直接向かう経路である．このように，途中の地点 B' から地点 C に向かう旅行者の最適戦略が，地点 A から地点 C に向かう旅行者の最適戦略の一部となっていることが，動的計画法が適用できる本質であり，これがベルマンの最適性原理で知られる性質である．

期待累積時間を最小とする最適戦略に従った場合，混雑時には 0.2 の確率で 160 分という長い時間を要してしまう．このような大きな損失を避けるのを目的とするのが，リスクを考慮した逐次的意思決定である．

大きな損失を避けるために条件付裾期待値 (CTE) を考える．CTE は条件付バリュー・アット・リスクとも呼ばれるリスク指標である．本稿の 2 節と 3 節では確率変数 X の信頼水準 α の CTE を

$$\mathsf{CTE}_\alpha[X] \equiv \mathsf{E}[X \mid X > V_\alpha] \tag{1}$$

と定義する．ただし，V_α は X の信頼水準 α のバリュー・アット・リスクとする．CTE に凸性などの性質を持たせるにはこの定義では不十分であるが，可読性のためにこの単純化した定義を利用する．また常に $\alpha = 0.8$ の信頼水準を考える．

すべての可能な戦略について累積時間 X の CTE を求めてみると，交通状況に依らずに，地点 A–地点 B'–地点 C の経路を通る戦略が最適であることがわかる．このとき，$\mathsf{CTE}_\alpha[X] = 96$ 分となる．また，前述の期待累積時間を最小にする戦略も，$\mathsf{CTE}_\alpha[X] = 96$ 分を達成し，同様に最適であることがわかる．

これでは前述の大きな損失を避けることができていないが，まずは動的計画法が適用できるかどうかを調べてみる．混雑時に地点 B' を出発して地点 C に向かう旅行者を考える．この旅行者の所要時間を Y とする．この旅行者にとって，$\mathsf{CTE}_\alpha[Y]$ を最小にするのは，地点 B' から地点 B を経由して地点 C に到達する経路である．これは，地点 A を出発する旅行者にとって $\mathsf{CTE}_\alpha[X]$ を最小とする 2 つの最適戦略のいずれとも相容れない戦略である．これはベルマンの最適性原理が成り立たないことを示しており，動的計画法によって最適戦略を算出することができない．

ベルマンの最適性原理が成り立たないという性質は，単純に最適戦略を動的計画法によって効率的に算出できないだけではなく，逐次的意思決定における目的関数として本質的に問題があるともいえる．地点 A から出発して $\mathsf{CTE}_\alpha[X]$ を最小とする最適戦略に従う旅行者は，10 分後に地点 B' に到着する．地点 B' において，混雑していることがわかったとする．混雑していることがわかった時点で，地点 A から出発した旅行者と，混雑しているという知識を持って地点 B' を出発する旅行者は，本質的に違いがないように思われる．前者は既に 10 分間の移動をしているが，地点 B' から同じ戦略に従った場合の両者の旅行時間は $X = Y + 10$ の関係にあり，CTE の性質から，$\mathsf{CTE}_\alpha[X] = \mathsf{CTE}_\alpha[Y] + 10$ が示せる．よって，$\mathsf{CTE}_\alpha[X]$ を最小とすることと $\mathsf{CTE}_\alpha[Y]$ を最小とすることは等価であるはずである．ベルマンの最適性原理が成り立たないということは，このような不整合が生じるということである．地点 A から地点 B' を経由して地点 C に向かう旅行者と，地点 B' から地点 C に向かう旅行者とでは，地点 B' からの最適戦略が異なる．

2 人の旅行者にとって，本質的に異なるのは，出発時点における交通状況に関する知識である．地点 A から出発する旅行者は，交通状況を知らずに最適戦略を立てる．そのとき，地点 B' から地点 C への経路が 150 分かかる確率は $0.2 \times 0.2 = 0.04$ と小さく，$\mathsf{CTE}_\alpha[X]$ の値には大きく寄与しない．これにより混雑時におい

ても地点 B' から地点 C に直接向かう戦略が最適戦略となる．地点 B' から交通状況の知識を持って出発する旅行者にとっては，混雑時には 150 分かかる確率は 0.2 と大きく，混雑時には地点 B を経由して地点 C に向かうことになる．

以上の議論を鑑みると，地点 B' から出発する旅行者のほうが，より良い意思決定をしているようにも考えられる．これは交通状況の知識を持って意思決定をできるからである．したがって，地点 A から出発した旅行者も地点 B' からは，地点 B' から出発する旅行者の最適戦略に従って旅行をすることも考えられる．

そのような，最適戦略を逐次更新していく旅行者は，交通状況に依らず，地点 A から地点 B' と地点 B をこの順に経由して地点 C に到達する．この経路は通常時で 65 分，混雑時で 125 分を要するが，いずれの場合も，地点 A から地点 B を経由して地点 C に到達する経路よりも 5 分長く要してしまう．

3. 反復的リスク指標に基づく逐次的意思決定

ここからはベルマンの最適性原理が成立するリスク指標を紹介していく．そのようなリスク指標の代表は，期待指数効用がある [3]．期待指数効用は確率変数 X をリスク感度 γ を用いて E[exp(γ X)] で実数値に写像する．期待指数効用で表現できるリスク選好は限定的であり [4]，本節では反復的リスク指標という広いクラスのリスク指標を考える．特に，反復的条件付バリュー・アット・リスク (ICTE) [5] を例に解説する．ICTE も CTE と同様に信頼水準 α をパラメタに持つが，常に $\alpha = 0.8$ を考える．

図 1 の例において，累積旅行時間の ICTE を最小にしようとする旅行者を考える．ある経路選択戦略にしたがったときの旅行時間 X の ICTE は反復的に CTE を計算することで算出できる．具体的には，まず地点 B か地点 B' に最初に到達して交通状況がわかったときの，旅行時間 X の CTE を $\mathsf{CTE}_\alpha[X|\Psi]$ として算出しておく．ここで Ψ は交通状況を表す確率変数である．出発時点においては，Ψ は確率変数であるから，$\mathsf{CTE}_\alpha[X|\Psi]$ も Ψ に応じた確率変数となる．さらに，この確率変数を CTE を用いて $\mathsf{CTE}_\alpha[\mathsf{CTE}_\alpha[X|\Psi]]$ のように実数値に写像することができる．この実数値を $\mathsf{ICTE}_\alpha[X]$ とする．

例えば，交通状況に依らずに地点 A から地点 B' を経由して地点 C に到達する経路を通る戦略の $\mathsf{ICTE}_\alpha[X]$ の値は以下のように計算される．まず，

$$\mathsf{CTE}_\alpha[X|\Psi = \text{混雑}]=160 \quad (2)$$
$$\mathsf{CTE}_\alpha[X|\Psi = \text{通常}]=70 \quad (3)$$

のように，交通状況がわかったとしたときの旅行時間の CTE を評価しておく．Ψ は 0.2 の確率で「混雑」，0.8 の確率で「通常」であるから，$\mathsf{CTE}_\alpha[X|\Psi]$ は 0.2 の確率で 160，0.8 の確率で 70 の値をとる確率変数である．したがって，$\mathsf{CTE}_\alpha[\mathsf{CTE}_\alpha[X|\Psi]] = 160$ と算出される．

実用的には，確率変数 Y と定数 c について $\mathsf{CTE}_\alpha[Y+c] = \mathsf{CTE}_\alpha[Y] + c$ となる CTE の性質を用いて，以下のように ICTE の値を計算できる．地点 B' からの所要時間 Y について，

$$\mathsf{CTE}_\alpha[Y|\Psi = \text{混雑}]=150 \quad (4)$$
$$\mathsf{CTE}_\alpha[Y|\Psi = \text{通常}]=70 \quad (5)$$

を計算しておく．次に，地点 A から地点 B' までの所要時間 10 分も考慮して，

$$\mathsf{CTE}_\alpha[\mathsf{CTE}_\alpha[X|\Psi]] = \mathsf{CTE}_\alpha[\mathsf{CTE}_\alpha[Y|\Psi] + 10] \quad (6)$$
$$= 160 \quad (7)$$

のように ICTE の値を算出する．このように ICTE の値は反復的に CTE を用いて計算される．上述の例は 2 期間であったが，これが多期間になれば，期間の数だけ CTE を反復的に適用させれば良い．ICTE は動的計画法による計算によって定義されていると考えることもできる．

ICTE の値を最小にする最適戦略も動的計画法によって求めることができる．それにはまず，Ψ が通常時と混雑時のそれぞれについて，地点 B からの $\mathsf{CTE}_\alpha[Y|\Psi]$ を最小にする戦略と地点 B' からの $\mathsf{CTE}_\alpha[Y|\Psi]$ を最小にする戦略を求める．地点 B からは直接地点 C に向かう経路しかないので，この経路を通るのが最適戦略となる．地点 B' からは 2 通りの経路があるが，交通状況に依存せずに，地点 B を経由して地点 C に到達する経路を通るのが最適な戦略である．これらの最適な戦略に従ったときの $\mathsf{CTE}_\alpha[Y|\Psi]$ の値は表 1 に示すとおりである．

地点 A からの最適な行動は，地点 B と地点 B' からの最適戦略に基づいて決定される．すなわち，地点 A

表 1 $\mathsf{CTE}_\alpha[Y|\Psi]$ の下限値

Ψ	通常	混雑
地点 B から	50	110
地点 B' から	55	115

から地点 B を経由して，地点 B から最適戦略に従って，地点 C に到達した場合と，地点 B を地点 B' に置き換えた場合を比較する．図 1 の例の場合，地点 A から地点 B や地点 B' までの所要時間は 10 分であるから，

$$\begin{aligned}&\min \mathsf{CTE}_\alpha[\mathsf{CTE}_\alpha[X|\Psi]]\\ =&\mathsf{CTE}_\alpha[\min \mathsf{CTE}_\alpha[Y|\Psi]+10]\end{aligned} \tag{8}$$

のような関係式が成り立つ．ただし，左辺の min は地点 A から可能な戦略についての最小値を表し，右辺の min は地点 B か地点 B' から可能な戦略の最小値を表す．上式は，CTE が単調性を持ち，確率順序 $Z_1 \geq Z_2$ について $\mathsf{CTE}_\alpha[Z_1] \geq \mathsf{CTE}_\alpha[Z_2]$ が成り立つことに起因する．これにより，地点 A から地点 B を経由して地点 C に至る経路を通る戦略が $\mathsf{ICTE}_\alpha[X]$ の値を最小（120 分）とする最適戦略であることがわかる．

図 1 の例においては $\mathsf{ICTE}_\alpha[X]$ を最小とする最適戦略に従うことで，160 分という長い時間を要する可能性を排除できている．これは動的計画法が適用できるというだけではなく，実際にリスクを考慮した逐次的意思決定が可能であることを示唆している．

多期間の場合にも，上述の 2 期間の逐次的意思決定の最適化と同様に，ICTE を最小とする戦略を動的計画法に基づいて決定することができる [4]．また，CTE を反復的に適用するのではなく，他のリスク指標を反復的に適用することによって，ICTE とは異なる反復的リスク指標が定義される．例えば，期待値と CTE の重み κ による凸結合によって，リスク指標

$$\rho(X) \equiv \kappa\, \mathsf{E}[X] + (1-\kappa)\, \mathsf{CTE}_\alpha[X] \tag{9}$$

が定義できる．このようなリスク指標を反復的に適用させて定義される反復的リスク指標を用いても良い [7]．

4. 反復的リスク指標の意味づけ

反復的リスク指標は，リスクを考慮した最適戦略の動的計画法による算出を可能にすることを見てきた．ところが，反復的リスク指標は再帰的に定義されているために，その意味が直感的に理解しにくいことがある．適切な意思決定のためには，目的関数の意味や性質を理解し，また目的関数のパラメタを適切に設定しておく必要がある．例えば，ICTE の値を最小にすることにどんな意味があるだろうか，また ICTE の持つ信頼水準 α はどのように設定すれば良いだろうか．本節では，反復的リスク指標，特に ICTE の意味の直感的な理解を深めることを目的とする．

まず $\mathsf{CTE}_\alpha[X]$ の意味を考える．前節までは $\mathsf{CTE}_\alpha[X] \equiv \mathsf{E}[X \mid X > V_\alpha]$ と単純化した定義を用いていたが，本節では以下の厳密な定義に従う：

$$\begin{aligned}&\mathsf{CTE}_\alpha[X]\\ &\equiv \frac{(1-\beta)\,\mathsf{E}[X \mid X > V_\alpha] + (\beta-\alpha)\,V_\alpha}{1-\alpha}.\end{aligned} \tag{10}$$

ただし，F_X を X の累積分布関数とし，$\beta = F_X(V_\alpha)$ と定義する．X が連続な確率分布を持つなど，X が値 V_α をとる確率が 0 の場合には，前節まで用いていた $\mathsf{CTE}_\alpha[X] \equiv \mathsf{E}[X|X > V_\alpha]$ の定義と一致する．

厳密な CTE の定義を用いると，以下の解釈が可能となる．まず，X を確率 p_i で値 v_i $(i = 1, \ldots, m)$ をとる確率変数とし，$v_1 > \ldots > v_m$ とする．このとき，$\mathsf{CTE}_\alpha[X]$ は以下の最適化問題の最適値と一致する：

$$\begin{aligned}\max_{\mathbf{q}} \quad & q_1\, v_1 + \ldots + q_m\, v_m\\ \text{s.t.} \quad & 0 \leq q_i \leq \frac{1}{\alpha} p_i, \forall i = 1, \ldots m\\ & q_1 + \ldots + q_m = 1.\end{aligned} \tag{11}$$

この最適化問題の解 $\mathbf{q} \equiv (q_1, \ldots, q_m)$ は，最適化問題の制約から確率の性質（$q_1 + \ldots + q_m = 1$ および $q_i \geq 0, i = 1, \ldots, m$）を満たす．また $\mathbf{q}$ が確率ベクトルであれば，この最適化問題の目的関数 $(q_1\, v_1 + \ldots + q_m\, v_m)$ は期待値と解釈できる．この期待値を最大にするような最悪ケースの $\mathbf{q}$ についての期待値が $\mathsf{CTE}_\alpha[X]$ となる．ここで制約条件に注目すると，$\mathbf{q}$ が確率の性質を満たす中で，各 q_i は p_i の $1/\alpha$ 倍まで大きな値をとることができる．目的関数値を大きくするには大きな v_i に対応する q_i をできるだけ大きな値に設定するのが良いので，最適解は図 2 に示すような $\mathbf{q}$ となることがわかる．すなわち，q_1 から順に，確率の性質を満たす範囲で，できるだけ大きな値に設定していき，確率ベクトル $\mathbf{p} \equiv (p_1, \ldots, p_m)$ の代わりに $\mathbf{q}$ を用いて期待値を計算する．直感的にもこれで $\mathsf{CTE}_\alpha[X]$ の値が計算できていることがわかるだろう．

このように考えると，$\mathsf{CTE}_\alpha[X]$ は確率に不確実性

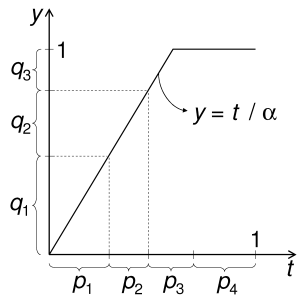

図 2 $\mathsf{CTE}_\alpha[X]$ を達成する最適解 $\mathbf{q}$ と $\mathbf{p}$ の関係

がある場合の，（値が小さいものほど良いとしたときの）最悪ケースの期待値として表現できることがわかる．具体的には，基準値が $\mathbf{p}$ によって与えられるが，真の確率ベクトル $\mathbf{q}$ は $0 \leq \mathbf{q} \leq \mathbf{p}/\alpha$ を満たす範囲にあることだけがわかっている場合である．

以上の議論を基に，MDP において，累積コストの ICTE を最小にすることの意味を考える．MDP は状態集合 S，行動集合 A，遷移確率関数 p，即時コスト関数 c の組 $\langle S, A, p, c\rangle$ で定義される．遷移確率関数は，状態 $s \in S$ において行動 $a \in A$ をとったときに，状態 $s' \in S$ に遷移する確率 $p(s'|s,a)$ を与える．ここでは，状態には期間に関する情報が付加されていると考える．このとき，状態集合 S は $S_0, \ldots, S_N$ と互いに疎な部分集合に分けられる．また，即時コスト関数は，状態 $s \in S$ において行動 $a \in A$ をとったときの即時報酬の分布を与えるとする．これらのパラメタがすべて厳密に既知である場合に，ある有限期間 N の累積コスト X に対して，$\mathsf{ICTE}_\alpha[X]$ を最小にする最適戦略を求める問題を議論してきた．戦略 π は，与えられた状態 $s \in S$ に対して，行動 $a \in A$ を決める関数である．

このようにリスクを考慮した逐次的意思決定は，パラメタに不確実性のある場合に，その最悪のケースに対して期待累積コストを最小とする逐次的意思決定 [6] と等価であることが示せる．特に，遷移確率関数が厳密にはわからないが，ある集合（不確実性集合）の中に入っていることはわかっている場合を考える．具体的には，前述の最適化問題のように，基準となる値 $p(s'|s,a)$ に対して，真の遷移確率 $q(s'|s,a)$ は $0 \leq q(s'|s,a) \leq p(s'|s,a)/\alpha$ を満たす範囲にあることがわかっているとする．また，確率であるので，各 $s \in S$, $a \in A$ について，$\sum_{s' \in S} q(s'|s,a) = 1$ の関係も満たす．このように遷移確率に不確実性がある場合には，遷移確率 q が不確実性集合 Q の中にある限り，期待累積コストがある一定値で抑えられるように，最悪ケースに対して最適な戦略 π を求めるロバストな最適化が考えられる．この遷移確率関数の不確実性に対してロバストに期待累積コストを最小化することと，遷移確率関数が既知である場合に累積コスト X の $\mathsf{ICTE}_\alpha[X]$ を最小化することとは等価である．すなわち，

$$\min_\pi \mathsf{ICTE}_\alpha[X^\pi] = \min_\pi \max_{q \in Q} \mathsf{E}^q[X^\pi] \qquad (12)$$

が成り立つ．ただし，X^π は戦略 π に従ったときの累積コストとする．また，左辺の ICTE_α は遷移確率 p を用いて計算されるとし，右辺の E^q は遷移確率 q を用いたときの期待値とする．

遷移確率関数に加えて，即時コスト関数に不確実性がある場合や，不確実性集合の形状が異なる場合についても，同様にロバストな逐次的意思決定が定義できることがある．また，これらのロバストな逐次的意思決定に対しても，パラメタ値が既知である場合にリスクを考慮した逐次的意思決定で等価なものが定義できることがある [7]．

本節では，累積コストの ICTE を最小化することの意味を，ロバストな逐次的意思決定として理解できることを紹介した．このようにリスクを考慮した逐次的意思決定とロバストな逐次的意思決定を対応づけることの意義は他に 2 点あると考えられる．まず，一方の逐次的意思決定だけを考えていると効率的な最適化アルゴリズムの設計が困難である場合に，他方の逐次的意思決定を考えることで，それが容易になることがある．次に，両者の対応づけを考えることによって，目的関数として用いる動機が得られることがある．これらの意義の例は [7] を参照されたい．

5. おわりに

本稿では，まず累積コストの CTE の最小化を目的とする逐次的意思決定には動的計画法が適用できず，最適な戦略が首尾一貫しないことをみた．このような最適戦略の一貫性は古くから議論されている．例えば，将来のコストを割り引く場合には，指数的に割り引かなければ首尾一貫した意思決定ができないことが知られている．また，首尾一貫しない意思決定者からは無限に搾取できることが知られている [8, pp. 30–31]．

次に，ICTE などの反復的リスク指標を用いることで，リスクを考慮した逐次的意思決定に動的計画法を適用できることを示し，その意味を議論した．CTE が確率変数を実数値に写像する関数（リスク指標）であるのに対して，ICTE は確率変数を確率変数の列に写像する動的リスク指標として定義するのが厳密である [5]．本稿ではこのような厳密性を犠牲にし，直感的な理解を優先させたことを注意されたい．本稿に関連するより厳密な議論は [4, 7] を参照されたい．

謝辞　本稿は独立行政法人科学技術振興機構 (JST), CREST の助成を受けて執筆されました．

参考文献

[1] Mausam and A. Kolobov, *Planning with Markov Decision Processes: An AI Perspective*, Morgan & Claypool Publishers, 2012.

[2] 恐神貴行，「これからのマルコフ決定過程」，日本オペレーションズ・リサーチ学会 2014 年春季シンポジウム予稿集，1–13, 2014.
[3] R. Howard and J. Matheson, “Risk-sensitive Markov decision processes,” *Management Science*, **18**, 356–369, 1972.
[4] T. Osogami, “Iterated risk measures for risk-sensitive Markov decision processes with discounted cost,” in *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, 567–574, 2011.
[5] M. R. Hardy and J. L. Wirch,“The iterated CTE: A dynamic risk measure,” *North American Actuarial Journal*, **8**, 62–75, 2004.
[6] A. Nilim and L. El Ghaoui, “Robust control of Markov decision processes with uncertain transition matrices,” *Operations Research*, **53**, 780–798, 2005.
[7] T. Osogami, “Robustness and risk-sensitivity in Markov decision processes,” *Advances in Neural Information Processing Systems*, **25**, 233–241, 2012.
[8] G. Ainslie, *Breakdown of Will*, Cambridge University Press, 2001.