マルコフ決定過程における Bellman 方程式の最適解探索アルゴリズムについて

On numerical algorithms for finding the solution to Bellman optimality equations in Controlled Markov Chains

* 源 隆哉 (Ryuya Minamoto), * 鈴木陸斗 (Rikuto Suzuki),
* 王瀚東 (Wang Handong), *,† 堀口正之 (Masayuki Horiguchi)

§1. はじめに

本研究では、多段決定過程としてマルコフ決定過程の最適化問題について考察し、評価基準が総割引期待利得のもとでの最適政策を求める最適解探索アルゴリズムとして、値反復法、政策反復法、修正政策反復法のそれぞれについて述べ、それぞれのアルゴリズムの具体的な計算例、アルゴリズムの特徴などについて、それらの結果をまとめる。まず、§2 では、マルコフ決定過程での基本的な構成モデルにおける用語の準備を行い、§3 では、総割引期待利得を目的関数とした Bellman 方程式と、最適政策の存在と一意性が保障される条件を明らかにする。§4 では、アルゴリズムを数値計算の観点から解析し、[7] に取り上げられている具体例を通じて、構成された探索アルゴリズムがどのように機能しているか、また、アルゴリズムの特徴について明らかにする。最後に §5 で、まとめと今後の課題について述べる。

* 神奈川大学大学院・理学研究科数学領域 (Field of Mathematics, Graduate School of Science, Kanagawa University),† 神奈川大学理学部理学科数学分野 (Department of Mathematics, Faculty of Science, Kanagawa University)

‡ 221-8686 横浜市神奈川区六角橋 3-27-1, (3-27-1, Rokkakubashi Kanagawa-ku, Yokohama City, Kanagawa Prefecture, 221-8686 Japan)

§2. 準備

総割引期待利得の最適政策を探す問題はマルコフ決定過程を用いて数理モデルとして定式化される。その際に必要な用語や記号などをここで準備する。

Definition 1 . Markov 決定過程の基本事項 ([1] 中出)

- S : **状態空間**または**状態集合**
  - 状態空間とは取りうる状態全体の集合である。基本的には状態に番号を振り、$S = \{0, 1, ...\}$ のように表現される。
- A : **決定空間**または**決定集合**
  - $s \in S$ のときに取りうる決定全体を $A(s)$ と書く。ただし $A := \bigcup_{s \in S} A(s)$
  - 時刻 t で状態 s での決定を $d_t(s)$ と書く。つまり $d_t(s) : S \times \mathbb{N} \cup \{0\} \to A(s)$
  - d_t と書いた場合は $d_t : S \to A$ とする。
- r : **利得関数**または**期待利得**
  - $s \in S$ で $a \in A(s)$ をとるとき、次の状態に推移するまでに受け取れる期待利得を $r(s, a)$ と書く。つまり $r : S \times A \to \mathbb{R}$
  - 最終時刻を T とするとき、これより先に推移することを考えることができないので、そのとき受け取れる期待利得は $K(s_T)$ と表す。
- p : **推移確率**
  - $s \in S$ で $a \in A(s)$ をとり状態 $s' \in S$ に推移する確率を $p(s'|s, a)$ と書く。
  - $p_{ij}^a = P(X_{n+1} = j | X_n = i, d_n(i) = a)$ で $d_n(i)$ は時刻 n で状態が i のときの決定を表す場合もある。
  - 状態 s_{t+1} は状態 s_t とそのときの決定 a_t のみで決まるので Markov 性を持つ。

確率過程 $\{X_t\}$ の確率空間によって、状態、決定、推移確率、期待利得の 4 つ組 $\{S, A, p, r\}$ の下での確率最適化モデルが構成される。これを、$\{S, A, p, r\}$ による Markov 決定過程と呼ぶことにする。状態空間については、S の構成要素によって、以下のように分類することができる。決定空間での分類についても同様である。

有限状態空間

有限個の状態から成る状態空間。有限状態空間に属する各状態に対して番号を付けることで状態空間を $S = \{0, 1, ..., N-1\}$ とすることができる。$(N < \infty)$

可算無限状態空間

可算無限個の状態から成る状態空間。可算無限状態空間に属する各状態に対して番号を付けることで状態空間を $S = \{0, 1, 2, ...\}$ とすることができる。

非可算無限状態空間

非可算無限個の状態から成る状態空間。状態空間が非可算無限個の要素から成るので番号付けすることができない。例えば実数で表現できる場合がある。

問題の目的としては今後出てくる**有限期間総期待利得**や**無限期間総期待利得**を最大にする決定の取り方を見つけることである。つまり、現在の状態を含む過去の状態と決定に関する**履歴**を基にして、各期において意思決定として最良の**政策**が得られることが目的である。そこで、以下のように履歴や政策を定義する。

Definition 2 . 履歴, 政策 ([1] 中出)

履歴

$H_t \subset S \times A(s_0) \times S \times \cdots \times A(s_{t-1}) \times S$ とする。このとき H_t を**時刻 t における履歴空間**といい、H_t 元を**時刻 t における履歴**という。$h_t := (s_0, a_0, ..., s_t)$ と書くことができる。

政策

政策とは、**決定の取り方のルール**を意味する。例えば、「時刻 t で履歴 h_t が与えられたとき、決定 a_t を確率 p_t でとる」といったものである。政策には、**定常性**（時刻に依存しない）、**Markov 性**（現在の状態のみに依存）、**決定性**（確率 1 で特定の決定をとる）などの性質がある。時刻依存の政策 π は、$\pi := (\pi_0, \pi_1, \dots)$ と表し、各 $\pi_t (t \in \mathbb{N} \cup 0)$ が決まれば政策が定まる。

これで用語などの準備は整った。以下は本資料で取り上げる問題の設定などを行う。

§3. 最適方程式

本節では、Markov 決定過程の評価基準のうちのひとつ、総期待利得基準の値関数の定式化を行い、続いて Bellman 方程式による値関数の特徴について概観する。

Definition 3. 総期待利得と Bellman 方程式 (中出 [1])

割引率

定数 $\gamma \in [0, 1)$ を**割引率**と呼ぶ。

有限期間総期待利得

$s_0 = s$, 政策 π, 期間 T **有限期間総期待利得**を以下の式で定義する。

$$V^{\pi}(s) := E^{\pi}\left[\sum_{t=0}^{T-1} \gamma^t r_t(s_t, a_t) + K(s_T) | s_0 = s\right]$$

π が定常政策のとき総期待利得 $V_{\gamma}^{\pi}(s)$ $(V^{\pi}(s))$ は、以下の方程式を満たす。

$$V_{\gamma}^{\pi}(s) = r(s, \pi(s)) + \gamma \sum_{s' \in S} p(s'|s, \pi(s)) V_{\gamma}^{\pi}(s')$$

総期待利得を最大化する政策 π を**最適政策**という。

Bellman 方程式

$V_{\gamma}^{*} := \sup_{\pi \in \Pi} V_{\gamma}^{\pi}(s)$ とする。本資料では max に到達させる π を見つけるのが目標だが次の関係式が成り立つ。

$$V_{\gamma}^{*}(s) = \max_{a \in A(s)} \left\{ r(s, a) + \gamma \sum_{s' \in S} p(s'|s, a) V_{\gamma}^{*}(s') \right\}$$

これを Bellman 方程式という。

この Bellman 方程式によって特徴づけられる総期待利得 $V_{\gamma}^{\pi}(s)$ $(V^{\pi}(s))$ の解が、最適値でありその方程式の解が最適政策である。マルコフ決定過程では、以下の最適化問題を解く。

最適化問題 (1)([1] 中出)

Bellman 方程式

$$V_{\gamma}^{*}(s) = \max_{a \in A(s)} \left\{ r(s, a) + \gamma \sum_{s' \in S} p(s'|s, a) V_{\gamma}^{*}(s') \right\}$$

の解 V_{γ}^{*} に対応する最適政策 π を求めよ。

ところで、この問題は動的計画法の**最適性の原理**から次のことが成り立つことがわかる。

最適化問題 (2)([1] 中出)

Bellman 方程式

$$V_{\gamma}^{*}(s) = \max_{a \in A(s)} \left\{ r(s, a) + \gamma \sum_{s' \in S} p(s'|s, a) V_{\gamma}^{*}(s') \right\}$$

の解 V_{γ}^{*} に対応する最適政策 π^* は以下のように記述される定常決定性 Markov 政策である。

$$\pi^*(s) := \arg \max_{a \in A(s)} \left\{ r(s, a) + \gamma \sum_{s' \in S} p(s'|s, a) V_{\gamma}^{*}(s') \right\}$$

上記の結果から $V^*(s)$ を求める必要がある。つまり Bellman 方程式を解いて $V^*_\gamma(s)$ を求める必要がある。以下、Bellman 方程式の解の存在、一意性に関して述べていく。

Definition 4 . Bellman 作用素
関数 u に対して以下のような作用素を定義する。

$$Lu(s) := \max_{a\in A(s)} \left\{ r(s,a) + \gamma \sum_{s'\in S} p(s'|s)u(s') \right\}$$

この作用素を **Bellman 作用素**という。Bellman 作用素に関しては以下が成り立つ。

Theorem 1 . Bellman 方程式の解の存在と一意性 ([[1] 中出, 定理 4.3,P.85])
$0 \leq \gamma < 1$ とし、$r(s,a)$ は一様有界とする。このとき、$V^*_\gamma(s)$ は Bellman 方程式 $V^\pi_\gamma(s) = LV^\pi_\gamma(s)$ の唯一解である。

最適解は、Bellman 方程式の解として得られることが、上記の定理によってわかるが、このときに問題になるのは Bellman 方程式の具体的な解き方である。Bellman 方程式を解いて V^*_γ を求めることができれば、その値を使って最適政策を求めることができる。そこで、次節において、Bellman 方程式の解の探索アルゴリズムについて考察する。

§4. Bellman 方程式の最適解探索アルゴリズム

Bellman 方程式の解き方、つまり最適政策を求める方法として**値反復法**と**政策反復法**が知られている。ここからは状態空間や決定空間は有限集合と仮定する。Bellman 作用素には縮小性があるので任意の関数 $V_0(s)$ に対して $V_n(s) = LV_{n-1}(s)$ とすると $n \to \infty$ のとき $\{V_n(s)\}$ は $V^*_\gamma(s)$ に収束する（Banach の不動点定理 ([3])）。このことから $V^*_\gamma(s)$ を求める方法として、次の値反復法を得る。

値反復法

1. 関数 $V_0(s), (s \in S)$ を選び, 初期値 $\varepsilon > 0$ を設定する。また、$n = 0$ とする。
2. 各 $s \in S$ に対して、$V_{n+1}(s)$ を次の式で計算する。
   $V_{n+1}(s) = \max_{a\in A(s)} \left\{ r(s,a) + \gamma \sum_{s'\in S} p(s'|s,a)V_n(s') \right\}$
3. $\|V_{n+1} - V_n\| = \max_{a\in A(s)} |V_{n+1}(s) - V_n(s)| \leq \varepsilon \dfrac{1-\gamma}{2\gamma}$ ならば 4 へ、そうでなければ n を 1 つ増やして 2 へ戻る。
4. 各 $s \in S$ について $\pi^*(s) = \arg\max_{a\in A(s)} \left\{ r(s,a) + \gamma \sum_{s'\in S} p(s'|s,a)V_{n+1}(s') \right\}$ として $\pi^*(s)$ を出力して終了。

この値反復法について、実際に、どのようにしてアルゴリズムの計算が進んでいくのかを、以下のように、具体例を通して見ていくことにする。

月次販売モデルの計算例（割引率 $\gamma = 0.9$ とする。）
(cf. Sheskin(2011), 5.2.1.2 Value Interation for Discounted MDP Model of Monthly Sales , P.374–)

再帰的状態からなる既約なマルコフ連鎖が、いずれの決定 a によって構成できるただ１つの communication class からなるマルコフ決定過程 $\{S, A, p, r\}$ について、各パラメータの具体的な数値は、以下の表 1 で与えられる月間販売モデルの最適化問題を考える。

状態　s	決定 a	p^a_{i1}	p^a_{i2}	p^a_{i3}	p^a_{i4}	$r(s,a)$
1(一番悪い状態)	1(メインではない資産の売却)	0.15	0.40	0.35	0.10	-30
	2(企業を買ってもらう)	0.45	0.05	0.20	0.30	-25
	3(従業員に早期退職を促す)	0.60	0.30	0.10	0	-20
2	1(経営陣の給与削減)	0.25	0.30	0.25	0.10	5
	2 (従業員の福利厚生の削減)	0.30	0.40	0.25	0.05	10
3	1(より良い製品の作成)	0.05	0.25	0.50	0.20	-10
	2 (新技術への投資)	0.05	0.25	0.50	0.20	-5
4 (一番いい状態)	1 (新しいプロジェクトへの投資)	0.05	0.20	0.40	0.35	35
	2 (戦略的な買収)	0	0.10	0.30	0.60	25

表 1　月間販売モデルでのマルコフ決定過程の構成要素 $\{S, A, p, r\}$

ここでは、7 ヶ月間の月次販売モデルを考察する。残り n 期で状態が s のときに得られる総期待利得を $V_n(s)$ と書くことにする。つまり $n = 0, 1..., 6$, $T = 7$、$V_7(s) = 0, (s = 1, 2, 3, 4)$ として以下のように定義する。
$V_n(s) = \max_{a \in A} \left\{ r(s,a) + \gamma \sum_{j=1}^{4} p^a_{sj} V_{n+1}(j) \right\}$ を表 1 の値を使って、各状態に対する最適な決定 (行動選択) を求めて、その後に最適政策を求めてく。

表 2 の初めの 1 列は時刻を表し、関数値 $V_n(s)$ は時刻 n, 状態 s における総割引期待利得、関数値 $d_n(s)$ は時刻 n, 状態 s における最適決定とする。また、$n = 6, 5, ..., 0$ の順で得られた $V_n(s), d_n(s)$ の値もそれぞれ示している。

	0	1	2	3	4	5	6	7
$V_n(1)$	-11.9208	-14.0600	-16.4959	-19.2459	-22.1155	-24.10	-20	0
$V_n(2)$	16.7625	14.7083	12.5184	10.3691	8.7276	8.65	10	0
$V_n(3)$	13.5505	11.5133	9.2951	6.8882	4.1643	0.40	-5	0
$V_n(4)$	61.5109	59.4047	56.9784	54.0470	50.254	45.125	35	0
$d_n(1)$	2	2	2	2	2	2	3	-
$d_n(2)$	2	2	2	2	2	2	2	-
$d_n(3)$	2	2	2	2	2	2	2	-
$d_n(4)$	2	2	2	2	2	1	1	-

表 2　V_n と d_n の値 (n=6,5,. . . ,1,0)

計算結果を見ると反復回数が 3 回超えたあたりから決定ベクトル $d = (2, 2, 2, 2)$ によって与えられる最適政策に収束したと見える。しかし、総割引期待利得の値に関しては収束しているかどうか判断は難しい。実際、政策反復法や線形計画法 (Linear Programming, LP) で解いた場合に求まる総割引期待利得の値と比較をしてみる。最初の状態を s としたときに得られる総割引期待利得を V_s と書くことにする。このとき、値反復法で求まった総割引期待利得の値は、順に、

$$V_1 = -11.9208, V_2 = 16.7625, V_3 = 13.5505, V_4 = 61.5109$$

である。他方、政策反復法や LP で、ある連立方程式を解いて求めた総割引期待利得の値は

$$V_1 = 6.8040, V_2 = 35.4613, V_3 = 32.2190, V_4 = 80.1970$$

となり、表 2 の値である値反復法での数値とは、かなり異なっていることがわかる。値が異なった理由として以下の理由が考えられる。

- 値反復法で求まる総割引期待利得の値は割引率 γ が $0 < \gamma < 1$ のとき Bellman 作用素が縮小写像になることを使って Banach の不動点定理を根拠に、Bellman 作用素をどんどん作用させて極限として求まるものであった。
- 政策反復法や LP で求まる総割引期待利得の値はある連立方程式を解いて求まる値であるので値反復法で求まる値と誤差がでてきる。
- さらに、今回は作用させた回数はたったの 7 回で最適政策は得られたものの、最終時点での V_n の値は近似総割引期待利得としては真の値には程遠く、本来の求めたい値は最適政策のもとでの真の総割引期待利得の値であるが、それらとの誤差が大きい数値例になっていた。

次に**政策反復法**について述べる。解くべき具体的な問題は、先ほどと同様に月間販売モデルとする。政策反復法はある連立方程式を解き続けて最適政策、総割引期待利得を求めていく。

政策反復法

1. 初期政策を任意に決定する。各状態 s に対して任意に決定 $d_0(s)$ を選ぶことで初期政策を任意に選ぶ。つまり初期政策として $d = (d_0(1), d_0(2), ..., d_0(N))$ を選ぶ。
2. 与えられた政策とそれに対応する p_{sj} と $r(s,a)$ を用いて、スタートの状態が s のときの総割引期待利得 V_s を求めるために

$$V_i = r(s, d_0(s)) + \gamma \sum_{j=1}^{N} p_{sj}^k V_j$$

   という連立方程式を解いて $V_1, ..., V_N$ を求める。
3. 各状態 i について、以下の値を最大にする決定 a^* を求める。

$$r(s,a) + \gamma \sum_{j=1}^{N} p_{sj}^a V_j$$

   状態空間、決定空間をそれぞれ有限空間であることを仮定しているので見つけることは可能。
4. 2 回連続する反復で全く同じ政策、すなわち任意の $s \in S$ に対して $d(s)$ が同じか前よりもいいものになれば停止する。つまり、全く更新されなくなったらアルゴリズムは終わる。

値反復法を試したときと同じモデルで政策反復法の具体的な計算を行う。まず、任意の初期政策を選ぶ。今回は $d = (3, 2, 2, 1)$ とする。利得行列は $q = (-20, 10, -5, 35)$ とする。このとき次の連立方程式を解いて V_1, V_2, V_3, V_4 を求める。$V_0(i) = V_i$ と書くことにする。このとき、連立方程式は、次のようになる。

$$\begin{cases} V_1 = r(1, d(1)) + \gamma(p_{11}V_1 + p_{12}V_2 + p_{13}V_3 + p_{14}V_4) \\ V_2 = r(2, d(2)) + \gamma(p_{21}V_1 + p_{22}V_2 + p_{23}V_3 + p_{24}V_4) \\ V_3 = r(3, d(3)) + \gamma(p_{31}V_1 + p_{32}V_2 + p_{33}V_3 + p_{34}V_4) \\ V_4 = r(4, d(4)) + \gamma(p_{41}V_1 + p_{42}V_2 + p_{43}V_3 + p_{44}V_4) \end{cases} \tag{1}$$

連立方程式 (1) を頑張って解くと、$V_1 = -38.2655, V_2 = 6.1707, V_3 = 8.1311, V_4 = 54.4759$

が得られる。次に、各状態 s に対して

$$V(s,a) := r(s,a) + \gamma(p_{i1}V_1 + p_{i2}V_2 + p_{i3}V_3 + p_{i4}V_4)$$

を最大にする a を求める。実際に計算してみると、次の表 3 の結果を得る。

決定 a	$V(1,a)$	$V(2,a)$	$V(3,a)$	$V(4,a)$
1	-25.4083	5.5205	-3.8312	54.4759
2	$\mathbf{-24.0478}$	$\mathbf{6.1707}$	$\mathbf{8.1311}$	$\mathbf{57.1677}$
3	-38.2655	-	-	-

表 3　政策反復法による計算例 (1)

よって、今回の数値例の問題では、得られる政策 (決定ベクトル) は $d = (2,2,2,2)$ となる。政策 $d = (2,2,2,2)$ に従って連立方程式 (1) を解き直せば、$V_1 = 6.8040$ $V_2 = 35.4613$ $V_3 = 32.2190$ $V_4 = 80.1970$ を得る。これらの値を使って再度、各状態 s に対して $V(s,a)$ 最大にする決定 a を求める。これを、計算をしてみる。

決定 a	$V(1,a)$	$V(2,a)$	$V(3,a)$	$V(4,a)$
1	1.0513	33.4722	21.9092	78.5501
2	$\mathbf{6.8040}$	$\mathbf{35.4613}$	$\mathbf{32.2190}$	$\mathbf{80.1970}$
3	-38.5158	-	-	-

表 4　政策反復法による計算例 (2)

よって、今回得られた政策 (決定ベクトル) は、前のステップと同様に $d = (2,2,2,2)$ でる。同じ政策を 2 回連続で得たのでここで政策反復法を終了して、最適政策として $d = (2,2,2,2)$、総割引期待利得として $V_1 = 6.8040$ $V_2 = 35.4613$ $V_3 = 32.2190$ $V_4 = 80.1970$ が得られる。値反復法と比較をしてみると、得られた最適政策は一致しているが、総割引期待利得は一致していないことがわかる。

最後に政策反復法の問題点と**修正政策反復法**について述べる。

政策反復法の問題点

1. Step 2 で連立方程式を解くが、状態空間が大きいと解くのにかなり時間がかかる。
2. 値反復法より収束は早いが、連立方程式を解くので 1 回の反復にかかる計算時間が長くなる。

などが挙げられる。そこで、次の修正政策反復法について述べる。これは政策を改良する際には $V_n(s)$ を正確には求めず、最適値関数の評価を逐次的に求めていく反復アルゴリズムである。

修正政策反復法

1. 初期関数 $V_0(s), \varepsilon > 0$, 自然数列 $a_n, n = 0, 1, 2, \cdots$ を決める。$n = 0$ とする。
2. $\pi_{n+1} \in \Pi_{m,d}$ を次の式で求める。

$$\pi_{n+1}(s) = \arg\max_{a \in A(s)} \left\{ r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) V_n(s') \right\}, s \in S$$

3. $k = 0$ とし、次の式で $w_{n,0}$ を求める。

$$w_{n,0} = \max_{a \in A(s)} \left\{ r(s,a) + \gamma \sum_{s' \in S} p(s'|s,a) V_n(s') \right\}, s \in S$$

4. $\|w_{n,0} - V_n\| \leq \varepsilon \dfrac{1-\gamma}{2\gamma}$ ならば 8 へ行く。
5. $k = a_n$ ならば 7 へ行く。そうでなければ次の式で $w_{n,k+1}$ を求める。

$$w_{n,k+1}(s) = r(s, \pi_{n+1}(s)) + \gamma \sum_{s' \in S} p(s'|s, \pi_{n+1}(s)) w_{n,k}(s'), s \in S$$

6. k を 1 増やす。5 に戻る。
7. $V_{n+1}(s) = w_{n,a_n}(s)$ とし、n を 1 つ増やし、2 に戻る。
8. π_{n+1} を近似最適政策として出力する。

これは V_γ^* を w_{n,a_n} で近似している。具体的な数値例については、省略とする。

§5. まとめ. 今後の課題

- 今回は値反復法、政策反復法、修正政策反復法による最適解探索アルゴリズムについて説明した。
- 今後の課題については状態空間や決定空間が抽象空間の場合などの抽象的な場合についての MDP の研究をしていきたい。
- 別の評価基準として平均利得マルコフ決定過程についても研究していきたい。
- 値反復法、政策反復法、修正政策反復法を実際に使う場合は、Python などでのプログラミングによる実装の必要がある。プログラミングやそれらの計算量について研究もしていきたい。

§. 参考文献

[1] 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

[2] 舟木直久. 確率論. 朝倉書店. 2022

[3] 竹内慎吾. 関数解析　基本と考え方. 裳華房. 2023

[4] 宮島静雄. 関数解析. 横浜図書. .2005

[5] 田中謙輔. 凸解析と最適化理論. オーム社. .2021

[6] M.L. Puterman. Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley. 1994.

[7] T.J. Sheskin. Markov Chains and Decision Process for Engineers and Managers. CRC Press. 2011