

相補的なポジティブ DP とネガティブ DP

船 木 洋 一

マルコフ決定過程において、割引総報酬が一定の値を超える確率を最大にする政策を求める問題は、はじめの状態を拡張して解くことができる。この帰着されたモデルは Positive Dynamic Programming (P D P) と呼ばれているモデルの一つである。これに対して、割引総報酬が最短時間で一定の値を超える政策を求める問題は、Negative Dynamic Programming (N D P) と呼ばれているモデルの一つに帰着される。これらのモデルは、一方のモデルの最適政策が他方のモデルにより、さらに改良されるという意味で補完的である。本稿ではその方法を示しそのことを明らかにする。

はじめに

システムの可能な状態の集合を S とする。システムの状態が i である時の可能な行動の集合を A_i とする。時点 t でシステムの状態が $i \in S$ のとき行動 $a \in A_i$ を選択すると、次の期に推移確率 p_{ij}^a で状態 j になりその期に直接報酬 r_{ij}^a が得られる。直接期待報酬は $r_i^a = \sum_j r_{ij}^a p_{ij}^a$ である。時点 t で行動を選択する規則を関数 d_t で表す。各時点での行動選択の規則の列 $\pi = \{d_0, d_1, \dots, d_n, \dots\}$ を政策と呼ぶ。 t 時点における状態を s_t 、行動を a_t とするとき $(s_0, a_0, s_1, a_1, \dots, a_{t-1}, s_t)$ を時点 t における履歴と呼ぶ。時点 t で行動を選択する規則が、履歴を考慮するかしないか、ランダムな選択を許すか許さないかによって、政策を以下のように呼ぶことにする。

マルコフ政策：現在の状態にのみ依存しランダムな選択を許す政策

定 常 政 策：マルコフ政策でランダムな選択を許さず、各時点の決定規則が同じ政策

割引率 (discount factor) を β ($0 < \beta < 1$) として、時点 t における直接期待報酬を $\tilde{r}(t)$ であらわすと、無限期間の割引総報酬 V は $V = \sum_t \beta^t \tilde{r}(t)$ であらわされる。初期状態 i と政策 π への依存を示すときにはそれぞれ $\tilde{r}_i^\pi(t)$ 、 V_i^π であらわす。目標は $V \geq C$ とする。状態 i で行動 a を選択して、報酬 r_{ij}^a を得て、次の時点で状態 j に推移したとする。そうすると、次の時点での目標は、 $V_j \geq (C - r_{ij}^a) / \beta$ となる。ただし V_j は次の時点以後の割引総報酬である。割引総報酬の超える目標、不等号の右辺を閾値という。次章以降では、もとのシステムの状態 i とこの目標の閾値 C をペアにした状態の推移を考える。

$$m = \min\{r_{ij}^a\}/(1-\beta), \quad M = \max\{r_{ij}^a\}/(1-\beta), \quad c_1 = m, \quad c_{n+1} = M$$

とする。 m を超えることは常に達成可能であるが M を超えることは常に達成不可能である。可能な閾値の集合を $\mathbf{K} = \{c_0, c_1, c_2, \dots, c_{n+1}\}$ と仮定する。これは閾値の最小単位がたとえば千である場合や、 M と m の間を n 等分した値のみを閾値にする場合などに相当する。 $c_{i+1} > c_i$ である。 $c_{k+1} > (C - r_{ij}^a)/\beta \geq c_k$ のとき $(C - r_{ij}^a)/\beta$ を c_k で近似する。 c_{n+1} 以上の時は c_{n+1} で近似する。 c_0 は十分小さな値で c_1 未満の値はすべて c_0 で近似する。本稿は論文[10]にあわせて小さい方の値で近似する。 \mathbf{K} は有限な集合である。簡単にするためさらに状態集合 $\mathbf{S} = \{1, 2, \dots, S\}$ も有限で、 \mathbf{A}_i も各 $i \in \mathbf{S}$ に対して有限であると仮定する。

[拡張]

$\mathbf{S} \times \mathbf{K}$ の要素を拡張された状態と呼び $i \in \mathbf{S}$ 、 $C \in \mathbf{K}$ として (i, C) で表す。また一つの文字でこの拡張された状態を表すときには x を用いることにする。行動空間は拡張された状態の1番目の要素のみに依存して、拡張前と同じである。

したがって、 $x = (i, C)$ のとき、 $\mathbf{A}_x = \mathbf{A}_i$ と定義する。

新しい状態を持った履歴を

$$(x_0, a_0, x_1, a_1, \dots, x_{t-1}, a_{t-1}, x_t)$$

$$x_k = (i_k, c_k) \in \mathbf{S} \times \mathbf{K}, \quad k = 0, 1, \dots, t; \quad a_k \in \mathbf{A}_{i_k}, \quad k = 0, 1, \dots, t-1$$

で定義し、この履歴を記号 h_t で表す。

また t 時点で履歴 h_t のときの行動 $a \in \mathbf{A}_{i_t}$ をとる確率が $q_{h_t}(a)$

$$\sum_a q_{h_t}(a) = 1, \quad 0 \leq q_{h_t}(a) \leq 1$$

であるランダムな決定関数を $d_t(h_t)$ であらわし、この決定関数からなる政策 $\{d_0, d_1, \dots\}$ を考える。

これ以降、単純に状態といったら拡張された状態 $x \in \mathbf{S} \times \mathbf{K}$ を指し、政策といったら拡張された履歴に依存する政策を指すことにする。マルコフ政策、定常政策の場合についてもそれらの依存する状態は $x \in \mathbf{S} \times \mathbf{K}$ とする。

記号の定義とPDP、NDP

$$\mathbf{U} = \{(i, C) \mid (i, C) \in \mathbf{S} \times \mathbf{K}, C \geq M\} \quad \mathbf{L} = \{(i, C) \mid (i, C) \in \mathbf{S} \times \mathbf{K}, C < m\}$$

$$\Delta_{(i,C)(j,D)}^a = \begin{cases} 1 & \text{if } D = (C - r_{ij}^a)/\beta \\ 0 & \text{if } D \neq (C - r_{ij}^a)/\beta \end{cases}$$

と定義する。 \mathbf{U} はどうしても達成不可能な集合であり、 \mathbf{L} は必ず達成可能な集合である。論文[10]定理2より

定理

拡張された状態の推移は、任意の政策に対して、確率 1 で **U** または **L** に入る。

が成立する。

また拡張された状態がひとたびこれら集合に入るとそこから外には出ない。Δ は、(j, D) が次に推移する状態である場合には 1、そうでない場合には 0 を与える記号である。

$x = (i, C)$, $y = (j, C')$, ただし $C' = (C - r_{ij}^a) / \beta$ として

$$\mathbf{U}_{x,y}^a = \begin{cases} 1 & \text{if } C' \geq M \\ 0 & \text{if } C' < M \end{cases} \quad \mathbf{L}_{x,y}^a = \begin{cases} 1 & \text{if } C' < m \\ 0 & \text{if } C' \geq m \end{cases}$$

と定義する。さらに

$$G_{(i,C)}^a = \sum_j \sum_D p_{ij}^a \cdot \Delta_{(i,C)(j,D)}^a \cdot \mathbf{U}_{(i,C),(j,D)}^a$$

$$R_{(i,C)}^a = \sum_j \sum_D p_{ij}^a \cdot \Delta_{(i,C)(j,D)}^a \cdot \mathbf{L}_{(i,C),(j,D)}^a$$

と定義する。G は次の期で **U** に入る確率を表し、R は **L** に入る確率を表す。

$$(G_{(i,C)}^a)_{(j,D)} = p_{ij}^a \cdot \Delta_{(i,C)(j,D)}^a \cdot \mathbf{U}_{(i,C),(j,D)}^a$$

$$(R_{(i,C)}^a)_{(j,D)} = p_{ij}^a \cdot \Delta_{(i,C)(j,D)}^a \cdot \mathbf{L}_{(i,C),(j,D)}^a$$

$$Q_{(i,C),(j,D)}^a = p_{ij}^a \cdot \Delta_{(i,C)(j,D)}^a \cdot (1 - \mathbf{U}_{(i,C),(j,D)}^a) \cdot (1 - \mathbf{L}_{(i,C),(j,D)}^a)$$

$$W_{(i,C)}^\pi = \Pr[V_i^\pi \geq C]$$

と定義する。Q は次の期で **U** にも **L** にも入らないで状態 (j, D) になる確率である。

[Positive DP]

$W_{(i,C)}^\pi = \Pr[V_i^\pi \geq C]$ は目的関数で、これを最大にする政策 π を見つけるのが PDP に帰着される一方の問題である。再帰関係式

$$W_x^\pi = \sum_a q(a) R_x^a + \sum_a q(a) \sum_y Q_{xy}^a W_y^\pi$$

が成り立つ。ただし π は π を 1 期ずらした政策である。

定常政策を $f^\infty = \{f, f, \dots, f, \dots\}$ と表す。 $f(x) = a$ の時に

$$G_x^f = G_x^a, \quad R_x^f = R_x^a, \quad Q_{xy}^f = Q_{xy}^a$$

と表す。定常政策 f^∞ のときには再帰関係式

$$W_x^{f^\infty} = R_x^f + \sum_y Q_{xy}^f W_y^{f^\infty}$$

が成り立つ。ここで $R_x^a = \sum_y (R_x^a)_y$ である。

目的関数： $\sup_{\pi} W_{(i,C)}^{\pi}$

状態空間： $\mathbf{S} \times \mathbf{K}$

行動空間： $\{\mathbf{A}_x : x \in \mathbf{S} \times \mathbf{K}\} \quad (x=(i, c) \text{ のとき } \mathbf{A}_x = \mathbf{A}_i)$

直接報酬： $\{R_x^a : x \in \mathbf{S} \times \mathbf{K}, a \in \mathbf{A}_i\}$

推移確率： $\{Q_{xy}^a : x \in \mathbf{S} \times \mathbf{K}, a \in \mathbf{A}_i\} \quad (\sum_y Q_{xy}^a \leq 1)$

は Positive DP (P D P)で、任意の (i, C) に対して $W_{(i,C)}^{\pi} = \Pr[V_i^{\pi} \geq C]$ を最大にする最適な定常政策が存在する ([5] [10] 参照)

[Negative DP]

t 時点までの総報酬を

$$V(t) = \tilde{r}(0) + \beta \tilde{r}(1) + \beta^2 \tilde{r}(2) + \cdots + \beta^t \tilde{r}(t)$$

と定義する。

$$Z(t) = 1 \quad \text{if} \quad V(t) < C$$

$$Z(t) = 0 \quad \text{if} \quad V(t) \geq C$$

と定義する。Z は総報酬が閾値を超えるか超えないかを見る関数である。

$$\tau = \sum_{t=1}^{\infty} Z(t), \quad T_x^{\pi} = E_x^{\pi}[\tau]$$

と定義する。 τ は総報酬が閾値を超えるまでの時間を表す。T は初期状態と政策を与えたときの総報酬が閾値を超える期待時間を表す。目標が達成されない確率が正のときは期待時間が ∞ となる。

$T_x^{\pi} = E_x^{\pi}[\tau]$ は目的関数であり、これを最小にする π を見つけるのが Negative DP に帰着される、もう一方の問題である。政策 π に対して再帰関係式

$$T_x^{\pi} = \sum_a q_x(a) \sum_y Q_{xy}^a T_y^{\pi} + \sum_a q_x(a) \sum_y G_{xy}^a \cdot \infty + \sum_a q_x(a) \sum_y Q_{xy}^a T_y^{\pi'}$$

成立する。

定常政策 f^{∞} のときには再帰関係式は

$$T_x^{f^{\infty}} = \sum_y \{Q_{xy}^f + (G_x^f)_y \cdot \infty\} + \sum_y Q_{xy}^f T_y^{f^{\infty}} \quad \dots (1)$$

となる。

目的関数： $\sup_{\pi} (-T_{(i,C)}^{\pi})$

状態空間： $\mathbf{S} \times \mathbf{K}$

行動空間： $\{\mathbf{A}_x : x \in \mathbf{S} \times \mathbf{K}\} \quad (x=(i, c) \text{ のとき } \mathbf{A}_x = \mathbf{A}_i)$

直接報酬： $\{-\sum_y \{(Q_x^a)_y + (G_x^a)_y \cdot \infty\} : x \in \mathbf{S} \times \mathbf{K}, a \in \mathbf{A}_i\}$

推移確率： $\{Q_{xy}^a : x \in \mathbf{S} \times \mathbf{K}, a \in \mathbf{A}_i\} \left(\sum_y Q_{xy}^a \leq 1 \right)$

とする。これはNegative DP(NDP)である。PDPと違いNDPでは直接報酬が有限とは限らない。したがって $(-T_{(i,C)}^\pi)$ も有限とは限らない。状態空間、行動空間が有限の場合には $\sup_\pi(-T_{(i,C)}^\pi)$ を満足する最適定常政策が存在する(Puterman [5] p312 Theorem7. 3. 6)。

両モデルの最適政策について

本稿のPDP、NDPの目的関数の基準をそれぞれ確率基準、時間最短基準と呼ぶことにしよう、

命題 1

確率基準最適政策によって、確率1で目的が達成可能な状態に対して、確率基準ではそれ以上最適政策を区別しない。しかしそれらの状態に対して、さらに時間最短基準で政策を改良することができる。

(証明) 確率1で目的が達成される状態に対しては、期待達成時間が有限である。これは状態の推移が確率1で集合Lに入って終わることを示している。推移の途中で、集合Uで終わる状態になることはない。したがって最後に集合Lで終わる状態の集合はそれ自身で閉じている。そこで、確率1で目的が達成される状態だけで $\sup_\pi(-T_{(i,C)}^{f^\infty})$ を目的関数として最適計算が可能であり、期待達成時間が最小の最適政策を求めることができる。

で確率基準で最適な定常政策 f^∞ により確率1で集合Lに入る拡張された状態の集合をあらわす。は閉じていて、 $x \in \Gamma$ に対して $W_x^{f^\infty} = 1$ であり、すべての y に対して $(G_x^f)_y = 0$ である。逆に $x \in \Gamma$ に対して、すべての y に対して $(G_x^h)_y = 0$ ならば式(1)より

$$-T_x^{h^\infty} = -\sum_{y \in \Gamma} Q_{xy}^h - \sum_{y \in \Gamma} Q_{xy}^h T_y^{h^\infty} \quad (x \in \Gamma)$$

であるから、 $T_x^{h^\infty}$ は有限となり、 $W_x^{h^\infty} = 1$ である。

はじめの確率基準で最適な政策 f^∞ に対して

$$-T_x^{f^\infty} = -\sum_{y \in \Gamma} Q_{xy}^f - \sum_{y \in \Gamma} Q_{xy}^f T_y^{f^\infty} \quad (x \in \Gamma)$$

が成立している。ここからスタートし時間最短基準で最適定常政策 g^∞ を見つけても

$$-T_x^{g^\infty} = -\sum_{y \in \Gamma} Q_{xy}^g - \sum_{y \in \Gamma} Q_{xy}^g T_y^{g^\infty} \quad (x \in \Gamma) \quad \dots (2)$$

が成立し、 $x \in \Gamma$ に対して $(G_x^g)_y = 0$ であり $W_x^{g^\infty} = 1$ である。従って g を集合L以外の状態では f と同じ決定をする関数として定義してやると、確率基準での最適性は損なわれず g^∞ は確率基準

でも最適な政策である。(証明終わり)

命題 2

時間最短基準で最適政策が見つかったとする。その最適政策を計算のスタートの政策として確率基準で最適政策を求めることにより、確率 1 で閾値を達成することができない状態に対して、閾値達成確率が最大となる政策を見つけることができる。その改良された最適政策は時間最短基準での最適性を損なわない。

(証明) 状態 $x \in \bar{\Gamma}$ ($\bar{\Gamma} = \mathbf{S} \times \mathbf{K} - \Gamma$) に対して時間最短基準 f^∞ で最適な政策からスタートして、確率基準最適政策を求める。

$$-T_x^{f^\infty} = -\infty \quad (x \in \bar{\Gamma})$$

$$W_x^{f^\infty} = R_x^f + \sum_{y \in \bar{\Gamma}} Q_{xy}^f W_y^{f^\infty} \quad (x \in \bar{\Gamma})$$

が成立している。 $\bar{\Gamma}$ から Γ への推移はあるが、いったん Γ に入ると $\bar{\Gamma}$ からでないの、吸収状態のあるマルコフ過程となり、 $\bar{\Gamma}$ に入る要素だけで計算が可能である。 $x \in \bar{\Gamma}$ に対して $W_x^{f^\infty} < 1$ であり、 Q_{xy}^f ($x \in \bar{\Gamma}, y \in \bar{\Gamma}$) は吸収状態のあるマルコフ連鎖であるから $R_x^f < 1$ となる。 $\bar{\Gamma}$ から Γ への推移が全くない場合はすべての $x \in \bar{\Gamma}$ に対して $W_x^{f^\infty} = 0$ であり、 $R_x^f = 0$ である。

最適解 g^∞ が得られたとして、

$$W_x^{g^\infty} = R_x^g + \sum_{y \in \bar{\Gamma}} Q_{xy}^g W_y^{g^\infty} \quad (x \in \bar{\Gamma}) \quad \dots (3)$$

が成立している。ある x に対して $W_x^{g^\infty} = 1$ が得られたとすると $0 \geq -T_x^{g^\infty} > -\infty = -T_x^{f^\infty}$ となり f^∞ が時間最短基準で最適であることに矛盾する。従ってすべての $x \in \bar{\Gamma}$ に対して $W_x^{g^\infty} < 1$ であり、状態 $x \in \Gamma$ に対しては g の行動の選択を f と同じとすると g^∞ は時間最短基準での最適性を損なわない。 g^∞ は時間最短基準では f^∞ と同じく最適であるが確率基準で見るとより良い政策である。(証明終わり)

一方の最適政策を他方の基準でさらに改良できるという意味で両基準は補完的である。しかし N D P では無限があらわれるので、単純に数値計算ができない。それで確率基準で最適政策を求め、その確率 1 で閾値が達成可能な状態に対して、時間最短基準で最適政策を求める、という方法で時間最短基準の最適政策を求めることができる。確率 1 で達成可能でない状態に対しては、確率基準で得られた最適政策と同じ行動を選択させる。

参考文献

- [1] Eitan Altman, *Constrained Markov Decision Processes*, (Chapman & Hall/CRC, 1999).
- [2] D. Blackwell, Positive dynamic programming, in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol.1 (University of California Press, Berkeley) pp. 415-418 (1967).
- [3] Lodewijk Kallenberg, Finite State and Action MDPs, in *Handbook of Markov Decision Processes Methods and Applications*, (Kluwer Academic Publishers) pp. 21-87 (2002).
- [4] H. Kushner, *Introduction to Stochastic Control Theory*, (Holt, Reinehart and Winston, New York, 1971).
- [5] M. L. Puterman, *Markov Decision Processes*, (John Wiley & Sons, New York, 1994).
- [6] Sheldon M. Ross, Dynamic Programming and Gambling Models, *Adv. Appl. Prob.* 6, 596-606 (1974)
- [7] R. Strauch, Negative dynamic programming, *Ann. Math. Stat.* 37, 871-890 (1966).
- [8] D. J. White, Minimizing a Threshold Probability in Discounted Markov Decision Processes, *J. Math. Anal. Appl.* 173, 634-646 (1993).
- [9] 船木洋一、 割引基準マルコフ決定理論 (東北大学博士論文) (1981) .
- [10] 船木洋一、 閾値確率基準マルコフ決定過程とポジティブDP、「人文社会論叢」(社会科学編)第11号、1 - 10(2004)