マルコフ決定過程における決定論的最適方策の存在証明

Gemini AI

2026 年 2 月 10 日

概要

この資料の査読（fact check）は実施していないので、参考程度にとどめよ。

1 序論

本稿では、割引無限ホライズン・マルコフ決定過程（Infinite Horizon Discounted MDP）において、決定論的（deterministic）な最適方策が少なくとも一つ存在することを証明する。証明の主軸として、ベルマン最適作用素がバナッハ空間上の縮小写像であることを利用し、バナッハの不動点定理を適用するアプローチをとる。

2 準備：定義と記法

定義 2.1 (マルコフ決定過程)**.** マルコフ決定過程（MDP）は、以下の 5 つ組 (S, A, P, R, γ) で定義される。

- S: 状態集合（有限集合とする）
- A: 行動集合（有限集合とする）
- $P : S \times A \times S \to [0, 1]$: 遷移確率関数。$P(s'|s, a)$ は状態 s で行動 a をとった時に状態 s' へ遷移する確率を表す。
- $R : S \times A \to \mathbb{R}$: 報酬関数（有界とする）。
- $\gamma \in [0, 1)$: 割引率。

定義 2.2 (価値関数とベルマンノルム)**.** 状態集合 S 上の実数値関数全体（価値関数の

集合）を $\mathcal{V} = \{v | v : S \to \mathbb{R}\}$ とする。$\mathcal{V}$ はベクトル空間であり、以下の最大値ノルム（sup-norm）を導入することでバナッハ空間（完備ノルム空間）となる。

$$\|v\|_\infty = \max_{s \in S} |v(s)| \tag{1}$$

定義 2.3 (ベルマン最適作用素)**.** 任意の価値関数 $v \in \mathcal{V}$ に対して、ベルマン最適作用素 $\mathcal{T}^* : \mathcal{V} \to \mathcal{V}$ を以下のように定義する。

$$(\mathcal{T}^* v)(s) = \max_{a \in A} \left[R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) v(s') \right] \tag{2}$$

3 主定理の証明

まず、ベルマン最適作用素が縮小写像であることを示す。

補題 3.1 (縮小写像性)**.** ベルマン最適作用素 $\mathcal{T}^*$ は、最大値ノルムに関して係数 γ の縮小写像である。すなわち、任意の $u, v \in \mathcal{V}$ に対して以下が成り立つ。

$$\|\mathcal{T}^* u - \mathcal{T}^* v\|_\infty \leq \gamma \|u - v\|_\infty \tag{3}$$

証明 任意の $s \in S$ を固定する。また、表記簡略化のため、$Q_v(s,a) = R(s,a) + \gamma \sum_{s'} P(s'|s,a) v(s')$ と置く。

$$\begin{aligned}
|(\mathcal{T}^* u)(s) - (\mathcal{T}^* v)(s)| &= \left| \max_{a \in A} Q_u(s,a) - \max_{a' \in A} Q_v(s,a') \right| \\
&\leq \max_{a \in A} |Q_u(s,a) - Q_v(s,a)| \quad (\because |\max f - \max g| \leq \max |f - g|) \\
&= \max_{a \in A} \left| \gamma \sum_{s' \in S} P(s'|s,a)(u(s') - v(s')) \right| \\
&= \gamma \max_{a \in A} \left| \sum_{s' \in S} P(s'|s,a)(u(s') - v(s')) \right| \\
&\leq \gamma \max_{a \in A} \sum_{s' \in S} P(s'|s,a) |u(s') - v(s')| \\
&\leq \gamma \max_{a \in A} \sum_{s' \in S} P(s'|s,a) \|u - v\|_\infty \\
&= \gamma \|u - v\|_\infty \sum_{s' \in S} P(s'|s,a) \\
&= \gamma \|u - v\|_\infty
\end{aligned}$$

これがすべての s で成り立つため、最大値ノルムをとることで $\|\mathcal{T}^* u - \mathcal{T}^* v\|_\infty \leq \gamma \|u - v\|_\infty$ が示された。 □

定理 3.2 (最適決定論的方策の存在)**.** 割引無限ホライズン *MDP* において、

1. 最適価値関数 V^* は一意に存在する。
2. V^* に対して貪欲な（*greedy*）決定論的方策 π^* は最適方策である。すなわち $V^{\pi^*} = V^*$。

証明　1. 最適価値関数の存在と一意性

$\mathcal{V}$ はバナッハ空間であり、補題 1 より $\mathcal{T}^*$ は縮小写像である。バナッハの不動点定理（Banach Fixed Point Theorem）により、$\mathcal{T}^* v = v$ を満たす不動点 $V^* \in \mathcal{V}$ がただ一つ存在する。これが最適価値関数である。

$$V^*(s) = \max_{a \in A} \left[R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^*(s') \right] \tag{4}$$

2. 決定論的方策の構成と最適性

各状態 s において、式 (4) の右辺を最大化する行動 a を選択する決定論的方策 π^* を構成する。

$$\pi^*(s) \in \operatorname*{argmax}_{a \in A} \left[R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) V^*(s') \right] \tag{5}$$

この方策 π^* に対するベルマン作用素 $\mathcal{T}^{\pi^*}$ を考えると、π^* の定義により、任意の s で以下の等式が成り立つ。

$$(\mathcal{T}^{\pi^*} V^*)(s) = R(s, \pi^*(s)) + \gamma \sum_{s'} P(s'|s, \pi^*(s)) V^*(s') = (\mathcal{T}^* V^*)(s) \tag{6}$$

V^* は $\mathcal{T}^*$ の不動点であるから、$(\mathcal{T}^* V^*) = V^*$ である。したがって、

$$\mathcal{T}^{\pi^*} V^* = V^* \tag{7}$$

となる。これは、V^* が方策 π^* のベルマン方程式の解であることを意味する。方策固定時のベルマン作用素 $\mathcal{T}^{\pi^*}$ もまた縮小写像であり、その不動点（すなわち方策 π^* の価値関数 V^{π^*}）は一意である。よって、$V^{\pi^*} = V^*$ が成立する。

以上より、決定論的方策 π^* は最適価値関数 V^* を達成するため、π^* は最適方策である。

□

4　結論

本証明により、割引 MDP においては、確率的方策を考慮に入れたとしても、決定論的な最適方策を常に構成可能であることが示された。これは、強化学習アルゴリズム（例：Q 学習）が決定論的な最適行動を探索することの理論的正当性を保証するものである。