数理解析研究所講究録
947巻 1996年 31-43

Continuous time Markov decision processes with discounted loss on a general state space

星野 満博 （新潟大学自然科学）

1 Introduction

本報告では、無限 (infinite horizon) 連続時間パラメータを有するマルコフ決定過程 (Markov decision processes) を扱う。連続時間マルコフ決定過程の研究としては、状態空間 (state space) が有限である場合を扱った Miller [7]、また Miller の結果を可算状態空間と可算行動空間 (countable action space) を伴う場合へ拡張した Kakumanu [5]、Qiying [8] などがある。[8] では推移確率の率 (transition probability rates) と利得率関数 (reward rate function) が時間に関して非定常 (nonstationary) である場合を扱っている。また、有限行動空間を伴う場合、最適ポリシー (optimal policies) が存在し、可算行動空間を伴う場合、任意の $\varepsilon > 0$ に対して、ε-最適ポリシー (ε-optimal policies) が存在することが知られている。状態空間と行動空間が一般の空間である場合、Doshi [2] が、あるポリシーが最適であるための必要十分条件とε-最適ポリシーが存在するための十分条件を与えている。このようなモデルにおいて、一般に最適ポリシーの存在性を示すためには行動空間のコンパクト性などのより強い条件が必要となる。

また、このような最適化問題を考える上で重要な役割を果たすのが最適方程式である。本報告では、状態空間と行動空間が一般空間の場合を扱い、[2] と異なるアプローチで、最適方程式、その解の存在性、更に ε-最適ポリシーの存在性について考察する。

2 Formulation and basic assumptions for the Markov decision system

本報告では、次の７個の対象物から成るマルコフ決定過程を考える 。

$$(\mathfrak{X}, \mathcal{A}, \mathcal{T}, r, \Pi, p_\pi, \alpha)$$

ここで、次のように仮定する。

(a) $\mathfrak{X}$ は動的決定システムの状態空間で、Polish（即ち、完備可分距離）空間の空集合でないボレル部分集合とする。$\mathcal{B}_\mathfrak{X}$ を $\mathfrak{X}$ のボレルσ-集合体とする。

(b) $\mathcal{A}$ は動的決定システムの行動空間で、Polish 空間の空集合でないボレル部分集合とする。$\mathcal{B}_{\mathcal{A}}$ を $\mathcal{A}$ のボレルσ-集合体とする。

(c) $\mathcal{T} = [0, \infty)$ は時刻集合。行動は、時間に関して連続的にとるものとする。

(d) $r : [0, \infty) \times \mathfrak{X} \times \mathcal{A} \to \mathbb{R}$ は損失率関数 (loss rate function) で、有界かつ可測とする。また、ある定数 M が存在して

$$r(t, x, a) \leq M, \quad \forall t, x, a$$

であると仮定する。

(e) Π は許容ポリシー (admissible policies) の集合で、(確率的) マルコフポリシーから成る。

(f) p_π はポリシー $\pi \in \Pi$ を使用したときの推移確率関数 (nonstationary transition probability function) である。すなわち、時刻 $s \geq 0$ において動的決定システムが状態 $x \in \mathfrak{X}$ であるとき、時刻 $t(\geq s)$ におけるシステムの状態は、確率測度 $p_\pi(s, x\,;\, t, \cdot)$ に従い決定される。推移確率関数 p_π は次の条件を満たすものとする。

(i) 各 $t \geq s \geq 0$、$x \in \mathfrak{X}$ に対して、$p_\pi(s, x\,;\, t, \cdot)$ は $\mathfrak{X}$ 上の確率測度で、

$$p_\pi(s, x\,;\, s, \{x\}) = 1 \quad \forall s \geq 0, \forall x \in \mathfrak{X}$$

である。

(ii) 各 $t \geq s \geq 0$、$\Gamma \in \mathcal{B}_{\mathfrak{X}}$ に対して、$p_\pi(s, \cdot\,;\, t, \Gamma)$ は $\mathfrak{X}$ 上の可測関数である。

(iii) Chapman-Kolmogorov の方程式

$$p_\pi(s, x\,;\, u, \Gamma) = \int_{\mathfrak{X}} p_\pi(t, y\,;\, u, \Gamma) p_\pi(s, x\,;\, t, dy), \qquad 0 \leq s \leq t \leq u.$$

が成立する。

(g) α は損失に関する割引率 (discount rate) で、正の定数とする。

本報告では、マルコフポリシー (Markov policies) だけに制限する。ここで、マルコフポリシー π は、$\mathcal{T} \times \mathfrak{X}$ が与えられたときの $\mathcal{A}$ 上のボレル可測確率核 (Borel measurable stochastic kernel) である。すなわち、

$\pi(\cdot | t, x)$: 各 $(t, x) \in \mathcal{T} \times \mathfrak{X}$ に対して、$\mathcal{A}$ 上の確率測度である。

$\pi(A | \cdot, \cdot)$: 各ボレル集合 $A \in \mathcal{B}_{\mathcal{A}}$ に対して、$\mathcal{T} \times \mathfrak{X}$ 上の可測関数である。

マルコフポリシー全体から成る集合を Π_M とする。各 $\pi(\cdot|t,x)$ が非確率的であるとき、マルコフポリシー π は確定的 (deterministic) であるという。多くの現実的応用モデルでは、様々な理由により使用できるポリシーはΠ_M の部分集合に制限されるであろう。それを Π と表し許容ポリシー (admissible policies) の集合と呼ぶことにする。

各ポリシー $\pi \in \Pi$ に対して、次の性質を持つ、確率過程 $\{X_t; t \geq 0\}$ が存在するものとする。

- $\{X_t; t \geq 0\}$ は強マルコフ過程。
- $\{X_t; t \geq 0\}$ のほとんど全ての sample path は、右連続かつ左側極限を持ち、任意の有限時間区間において、不連続点は高々有限個である。

次のような最適基準を考える。システムが時刻 $t \geq 0$、状態 $x \in \mathfrak{X}$ でスタートし、ポリシー $\pi \in \Pi$ を使用した場合、総期待割引損失 (total expected discounted loss) を以下のように与える。

$$V_\pi^\alpha(t,x) = E_\pi\left[\int_t^\infty e^{-\alpha(s-t)} r_\pi(s, X_s) ds \,\middle|\, X_t = x\right] \tag{2.1}$$

ただし、E_π は推移確率関数 p_π に関する期待値作用素 (expectation operator) で、$e^{-\alpha} \in (0,1)$ は割引因子 (discount factor) を意味する。また $r_\pi : \mathcal{T} \times \mathfrak{X} \to \mathbb{R}$ は損失率関数 (loss rate function) で

$$r_\pi(t,x) = \int_{\mathcal{A}} r(t,x,a)\pi(da|t,x), \qquad t \geq 0,\ x \in \mathfrak{X}$$

によって定義する。r_π の有界性から、明らかに

$$|V_\pi^\alpha(t,x)| \leq \frac{M}{\alpha}, \quad \forall \pi \in \Pi$$

が成り立つ。(2.1) において期待値作用素と積分の交換を仮定することにより

$$\begin{aligned} V_\pi^\alpha(t,x) &= \int_t^\infty e^{-\alpha(s-t)} E_\pi[r_\pi(s, X_s)|X_t = x] ds \\ &= \int_0^\infty e^{-\alpha s} \left\{ \int_{\mathfrak{X}} r_\pi(t+s, y) p_\pi(t, x\,;\, t+s, dy) \right\} ds \end{aligned} \tag{2.2}$$

を得る。

このとき、動的決定システムにおいて、次の最小化問題 (P) を考える。

(P) Minimize $V_\pi^\alpha(t,x)$ subject to $\pi \in \Pi$

問題 (P) に対して、我々の最終的な目的は最適ポリシーまたは ε-最適ポリシーを求めることである。ここで、最適ポリシー、ε-最適ポリシーを以下のように定める。

Definition 2.1 (i) 最適値関数 (optimal value function) を

$$V^{\alpha}_{\mathrm{opt}}(t,x) = \inf_{\pi\in\Pi} V^{\alpha}_{\pi}(t,x)$$

によって定義する。

(ii) 次の等式が成り立つとき、$\pi^* \in \Pi$ は Π において最適 (optimal) であるという。

$$V^{\alpha}_{\pi^*}(t,x) = \inf_{\pi\in\Pi} V^{\alpha}_{\pi}(t,x), \qquad \forall (t,x) \in \mathcal{T}\times\mathfrak{X}$$

(iii) 与えられた定数 $\varepsilon > 0$ に対して、次の不等式が成り立つとき、$\pi_\varepsilon \in \Pi$ はΠ において ε-最適 (ε-optimal) であるという。

$$V^{\alpha}_{\pi_\varepsilon}(t,x) \le V^{\alpha}_{\mathrm{opt}}(t,x) + \varepsilon, \qquad \forall (t,x) \in \mathcal{T}\times\mathfrak{X}$$

3 Optimality equation and existence of ε-optimal policies

最適方程式 (optimality equation) を導くための準備として、最初に次のような関数の集合を与える。$\mathcal{T}$ の部分集合 $\mathcal{T}'$ を時間区間とする。

- $I(\mathcal{T}'\times\mathfrak{X})$ ：各 $t \in \mathcal{T}'$ に対して、$u(t,\cdot)$ が $\mathfrak{X}$ 上の有界かつ可測な実数値関数であるような関数 $u : \mathcal{T}'\times\mathfrak{X} \to \mathbb{R}$ の全体。
- $B(\mathcal{T}'\times\mathfrak{X})$ ：$\mathcal{T}'\times\mathfrak{X}$ 上の有界かつ可測な実数値関数 $u : \mathcal{T}'\times\mathfrak{X} \to \mathbb{R}$ の全体から成るバナッハ空間。ここで、ノルムは supremum norm

$$\|u\| = \sup_{(t,x)\in\mathcal{T}'\times\mathfrak{X}} |u(t,x)|$$

とする。

- $\mathcal{D}(\mathcal{T}'\times\mathfrak{X})$ ：次の条件を満たす $\mathcal{T}'\times\mathfrak{X}$ 上の有界かつ可測な実数値関数 u の全体。
  - (i) 任意の x に対して、関数 $u(\cdot,x)$ の各点 $t \in \mathcal{T}'$ における微分係数 $D_t u(t,x)$ が存在がする。
  - (ii) 関数 $D_t u(\cdot,x)$、$D_t u(t,\cdot)$ は共に可測で、さらに $|D_t u(t,x)| \le b(t)$ を満たす連続関数 b が存在する。

各 $\pi \in \Pi$、$s \ge 0$ に対して、次のように定義される線形作用素 $P^{\pi}_s : I(\mathcal{T}\times\mathfrak{X}) \to I(\mathcal{T}\times\mathfrak{X})$ を導入する。

$$\begin{aligned} P^{\pi}_s u(t,x) &= E_\pi\left[u(X_{t+s})|X_t = x\right] \\ &= \int_{\mathfrak{X}} u(t+s,y) p_\pi(t,x\,;\,t+s,dy), \qquad (t,x) \in \mathcal{T}\times\mathfrak{X} \end{aligned} \tag{3.1}$$

このとき、推移確率関数の特性により、作用素 P_s は次の性質をもつ。

(i) $P_0^\pi = I$ (identity operator)

(ii) $P_{s_1}^\pi P_{s_2}^\pi = P_{s_1+s_2}^\pi, \quad \forall s_1, s_2 \geq 0$

(iii) $\|P_s^\pi u\| \leq \|u\|, \quad \forall u \in B(\mathcal{T}\times\mathfrak{X})$ (by the Chapman-Kolmogorov equation)

また、(2.2) から等式

$$V_\pi^\alpha(t,x) = \int_0^\infty e^{-\alpha s} P_s^\pi r_\pi(t,x)ds \tag{3.2}$$

が成り立つ。

まず最初に、各推移確率関数 p_π に対して次の仮定を課す。

Assumption A (i) 全ての $(t,x) \in \mathcal{T}\times\mathfrak{X}$、$\Gamma \in \mathcal{B}_\mathfrak{X}$ に対して、関数 $p_\pi(t,x\,;\,\cdot,\Gamma)$ は微分可能である。

(ii) 次の条件を満たす関数 $q_\pi : \mathcal{T}\times\mathfrak{X}\times\mathcal{B}_\mathfrak{X} \to \mathbb{R}$ が存在する。

(a) 各 $q_\pi(t,x\,;\,\cdot)$ は $\mathfrak{X}$ 上の符号付測度 (signed measure) で、各 $q_\pi(t,\cdot\,;\,\Gamma)$ は可測関数。

(b) ある連続関数 c に対して、$q_\pi(t,x\,;\,\{x\}) \geq -c(t)$

(c) 任意の $s \geq t$ に対して、

$$D_s p_\pi(t,x\,;\,s,\Gamma) = \int_\mathfrak{X} q_\pi(s,y\,;\,\Gamma)p_\pi(t,x\,;\,s,dy) \tag{3.3}$$

Lemma 3.1 (i) $x \in \Gamma$ であるならば、$-c(t) \leq q_\pi(t,x\,;\,\Gamma) \leq 0$

(ii) $x \notin \Gamma$ であるならば、$0 \leq q_\pi(t,x\,;\,\Gamma) \leq c(t)$

(iii) $q_\pi(t,x\,;\,\mathfrak{X}) = 0$

(iv) $D_s p_\pi(t,x\,;\,s,\cdot)$ は $\mathfrak{X}$ 上の符号付測度である。

Proof. (3.3) において $s = t$ とすると

$$q_\pi(t,x\,;\,\Gamma) = \lim_{h\downarrow 0}\frac{1}{h}\Big\{p_\pi(t,x\,;\,t+h,\Gamma) - \delta_x(\Gamma)\Big\} \tag{3.4}$$

が得られる。ここで、δ_x は δ 測度で、$x \in \Gamma$ のとき $\delta_x(\Gamma) = 1$、$x \notin \Gamma$ のとき $\delta_x(\Gamma) = 0$ である。(3.4) から、性質 (iii) が導かれ、また $x \in \Gamma$ のとき $q_\pi(t,x\,;\,\Gamma) \leq 0$、$x \notin \Gamma$ のとき $q_\pi(t,x\,;\,\Gamma) \geq 0$ となる。従って、$x \in \Gamma$ のとき

$$q_\pi(t,x\,;\,\Gamma) \geq q_\pi(t,x\,;\,\{x\}) \geq -c(t)$$

が成り立ち、性質 (i) を得る。さらに性質 (iii) により $q_\pi(t,x\,;\,\Gamma) = -q_\pi(t,x\,;\,\mathfrak{X}\backslash\Gamma)$ が導かれるので、性質 (i) から $x \notin \Gamma$ のとき $q_\pi(t,x\,;\,\Gamma) \leq c(t)$ が成り立ち、性質 (ii) を得る。

次に性質 (iv) について、(3.3) より、任意の互いに素である$\Gamma_n \in \mathcal{B}_{\mathfrak{X}}$、$n = 1, 2, \cdots$ に対して、

$$D_s p_\pi(t,x\,;\,s, \cup_n \Gamma_n) = \int_{\mathfrak{X}} \sum_n q_\pi(s,y\,;\,\Gamma_n) p_\pi(t,x\,;\,s,dy)$$

であり、ここで、性質 (i)、(ii) から

$$\left|\sum_{n=1}^{m} q_\pi(s,y\,;\,\Gamma_n)\right| = \left|q_\pi(s,y\,;\,\cup_{n=1}^{m}\Gamma_n)\right| \le c(s)$$

が成り立つので、積分記号と総和記号の入れ替えが可能となり、$D_s p_\pi(t,x\,;\,s,\cdot)$ はσ 加法性を持つ。よって、符号付測度である。 □

A_t, $t \ge 0$ を p_π の生成作用素 (infinitesimal generators) とするとき、ボレル集合 Γ と定義関数 $I_\Gamma(x)$ に対して $A_t I_\Gamma$ が存在するならば $q_\pi(t,x\,;\,\Gamma) = A_t I_\Gamma(x)$ が成り立つ。非定常推移確率関数の生成作用素の定義については Friedman [4, Chapter 2] を参照されたし。

$B(\mathcal{T}\times\mathfrak{X})$ 上の線形作用素 Q_π を

$$\begin{aligned} Q_\pi u(t,x) &= \int_{\mathfrak{X}} u(t,y) q_\pi(t,x\,;\,dy) \\ &= \int_{\mathfrak{X}} u(t,y) q_\pi^+(t,x\,;\,dy) - \int_{\mathfrak{X}} u(t,y) q_\pi^-(t,x\,;\,dy), \quad (t,x) \in \mathcal{T}\times\mathfrak{X} \end{aligned}$$

によって定義する。ここで、$q_\pi^+(t,x\,;\,\cdot)$, $q_\pi^-(t,x\,;\,\cdot)$ はそれぞれ符号付測度 $q_\pi(t,x\,;\,\cdot)$ のジョルダン分解 (Jordan decomposition) による正変分 (upper variation measures)、負変分 (lower variation measures) である。すなわち、任意の $\pi \in \Pi$、$u \in B(\mathcal{T}\times\mathfrak{X})$、$(t,x) \in \mathcal{T}\times\mathfrak{X}$ に対して、

$$Q_\pi u(t,x) = \int_{\mathfrak{X}\setminus\{x\}} u(t,y) q_\pi(t,x\,;\,dy) + u(t,x) q_\pi(t,x\,;\,\{x\})$$

である。さらに、Lemma 3.1(i), (ii) から

$$\begin{aligned} |Q_\pi u(t,x)| &\le \int_{\mathfrak{X}\setminus\{x\}} |u(t,y)|\, q_\pi(t,x\,;\,dy) + |u(t,x) q_\pi(t,x\,;\,\{x\})| \\ &\le q_\pi(t,x\,;\,\mathfrak{X}\setminus\{x\}) \sup_{x\in\mathfrak{X}} |u(t,x)| - q_\pi(t,x\,;\,\{x\}) \sup_{x\in\mathfrak{X}} |u(t,x)| \\ &\le 2c(t)\|u\| \end{aligned} \tag{3.5}$$

が成り立つので、$Q_\pi u \in I(\mathcal{T}\times\mathfrak{X})$ である。

Lemma 3.2 $\pi \in \Pi$, $V \in \mathcal{D}(\mathcal{T}\times\mathfrak{X})$, $g \in B(\mathcal{T}\times\mathfrak{X})$ とする。全ての $x \in \mathfrak{X}$ に対して、

$$\alpha V(t,x) \underset{(\ge)}{\le} r_\pi(t,x) + Q_\pi V(t,x) + D_t V(t,x) + g(t,x) \qquad \text{a.a.}\, t \in \mathcal{T}$$

であるならば、

$$V(t,x) \underset{(\ge)}{\le} V_\pi^\alpha(t,x) + \int_0^\infty e^{-\alpha s} P_s^\pi g(t,x) ds, \quad \forall (t,x) \in \mathcal{T}\times\mathfrak{X} \tag{3.6}$$

が成り立つ。ここで a.a. $t \in \mathcal{T}$ は、ほとんど全ての *(almost all)* $t \in \mathcal{T}$ を意味する。

Proof. 各 $t \in \mathcal{T}$ に対して、関数 $V(t,\cdot)$、$Q_\pi V(t,\cdot)$、$D_t V(t,\cdot)$ は可積分であるから、仮定と P_s^π の単調性から

$$\alpha e^{-\alpha s} P_s^\pi V - e^{-\alpha s} P_s^\pi Q_\pi V - e^{-\alpha s} P_s^\pi D_t V \underset{(\geq)}{\leq} e^{-\alpha s} P_s^\pi r_\pi + e^{-\alpha s} P_s^\pi g \qquad \text{a.a. } s \geq 0 \quad (3.7)$$

を得る。

次に、各 $(t,x) \in \mathcal{T}\times\mathfrak{X}$ に対して、関数 $F(s) := -e^{-\alpha s} P_s^\pi V(t,x)$ が $[0,\infty)$ 上、微分可能で、導関数 F' が不等式 (3.7) の左辺と等しいことを示すことにする。まず、0 のある近傍が存在して、これに属す全て h のに対して、

$$\begin{aligned}\frac{1}{h}\Big\{P_{s+h}^\pi V(t,x) - P_s^\pi V(t,x)\Big\} = \frac{1}{h}\int_{\mathfrak{X}}\Big\{V(t+s+h,y) - V(t+s,y)\Big\}p_\pi(t,x\,;\,t+s,dy)\\ + \int_{\mathfrak{X}}\int_{\mathfrak{X}} V(t+s+h,y)q_\pi(t+s,z\,;\,dy)p_\pi(t,x\,;\,t+s,dz)\\ + \frac{1}{h}\int_{\mathfrak{X}} V(t+s+h,y)\psi(h,dy) \qquad (3.8)\end{aligned}$$

が成り立つ。$V \in \mathcal{D}(\mathcal{T}\times\mathfrak{X})$ から、微分係数 $D_t V(t+s,x)$ が存在し、$\sup_{x\in\mathfrak{X}} |D_t V(t+s,x)| \leq b(t+s)$ である。従って、任意の $s \geq 0$ に対して、(3.8) の右辺の最初の積分は $h \to 0$ のとき $P_s^\pi D_t V(t,x)$ に収束する。更に、関数 $V(\cdot,x)$ の連続性と関数 V の有界性から、$h \to 0$ のとき、(3.8) の右辺の 2 番目の積分は $P_s^\pi Q_\pi V(t,x)$ へ、3 番目の積分は 0 へ収束する。その結果、(3.8) から全ての $(t,x) \in \mathcal{T}\times\mathfrak{X}$ に対して、

$$\lim_{h\to 0}\frac{1}{h}\Big\{P_{s+h}^\pi V(t,x) - P_s^\pi V(t,x)\Big\} = P_s^\pi Q_\pi V(t,x) + P_s^\pi D_t V(t,x) \qquad \forall s \geq 0$$

となる。従って、F は微分可能で

$$F'(s) = \alpha e^{-\alpha s} P_s^\pi V(t,x) - e^{-\alpha s}\frac{d}{ds} P_s^\pi V(t,x)$$

これらの式から、(3.7) の左辺は F' と等しく、

$$\frac{d}{ds}\left[-e^{-\alpha s} P_s^\pi V(t,x)\right] \underset{(\geq)}{\leq} e^{-\alpha s} P_s^\pi r_\pi(t,x) + e^{-\alpha s} P_s^\pi g(t,x) \qquad (3.9)$$

を得る。(3.9) の両辺を $s = 0$ から $s = \tau$ まで積分することにより、

$$V(t,x) - e^{-\alpha\tau} P_\tau^\pi V(t,x) \underset{(\geq)}{\leq} \int_0^\tau e^{-\alpha s} P_s^\pi r_\pi(t,x)ds + \int_0^\tau e^{-\alpha s} P_s^\pi g(t,x)ds \qquad \forall \tau > 0$$

が成り立つ。ここで、極限 $\tau \to +\infty$ をとり等式 (3.2) を用いることにより、不等式 (3.6) が得られる。 □

Lemma 3.2 において、$g = 0$ または $g = \alpha\varepsilon 1$ とおくことにより、直ちに、あるポリシーが最適または ε-最適であるための十分条件が得られる。

Theorem 3.1 $\pi^*, \pi^{**} \in \Pi$、 $V, V^\alpha_{\pi^*}, V^\alpha_{\pi^{**}} \in \mathcal{D}(\mathcal{T}\times\mathfrak{X})$ とする。

(i) 各 $\pi \in \Pi$ に対して、関数 V が方程式

$$\alpha V(t,x) = r_\pi(t,x) + Q_\pi V(t,x) + D_t V(t,x) \qquad \text{a.a.}\, t \in \mathcal{T},\ x \in \mathfrak{X} \tag{3.10}$$

の解であるならば、$V = V^\alpha_\pi$ が成り立つ。

(ii) 関数 $V^\alpha_{\pi^*}$ が方程式

$$\alpha V(t,x) = \inf_{\pi\in\Pi}\left\{r_\pi(t,x) + Q_\pi V(t,x) + D_t V(t,x)\right\} \qquad \text{a.a.}\, t \in \mathcal{T},\ x \in \mathfrak{X} \tag{3.11}$$

の解であるならば、ポリシー π^* は最適である。

(iii) $\varepsilon > 0$ を任意の定数とするとき、関数 $V^\alpha_{\pi^{**}}$ が方程式

$$\alpha V(t,x) = \inf_{\pi\in\Pi}\left\{r_\pi(t,x) + Q_\pi V(t,x) + D_t V(t,x)\right\} + \alpha\varepsilon \qquad \text{a.a.}\, t \in \mathcal{T},\ x \in \mathfrak{X}$$

の解であるならば、ポリシー π^{**} は ε-最適である。

Theorem 3.1(ii) の観点において、(3.11) を最適方程式とよぶことにする。最適方程式は我々の最小化問題 (P) を解析するうえで重要な役割を果たす。最適方程式の解の存在性を示すために幾つかの補題を準備することにしよう。

Lemma 3.3 任意の定数 $\lambda \in [0,1)$ と任意の有限区間で可積分である任意の関数 $f : \mathcal{T} \to [0,\infty)$ に対して、次の条件を満たす狭義単調増加実数列 $\{\tau_n\}_{n\geq 0}$ が存在する。

$$\tau_0 = 0, \quad \tau_n \uparrow +\infty, \quad \int_{\tau_n}^{\tau_{n+1}} f(s)ds \leq \lambda$$

Proof. 任意の $\lambda \in [0,1)$ と f に対する実数列 $\{\tau_n\}_{n\geq 0}$ の選び方は Qiying [8] を参照されたし。 □

Lemma 3.3 により、与えられた定数 $\lambda \in [0,1/2)$ と関数 $\alpha 1 + 2c$ に対して、Lemma 3.3 と同様な性質をもつ時刻列 $\{t_n\}_{n\geq 0}$ を取ることが出来る。つまり

$$t_0 = 0, \qquad t_n \uparrow +\infty, \qquad \int_{t_n}^{t_{n+1}} \{\alpha + 2c(s)\}ds \leq \lambda \quad \forall n \geq 0$$

ここで、関数 c は Assumption A(ii)(b) の連続関数である。この時刻列 $\{t_n\}$ に対して、$\mathcal{T}_n = [t_n, t_{n+1}]$ とおき、$B(\mathcal{T}_n\times\mathfrak{X})$ 上の作用素 S_n, $n \geq 0$ を

$$S_n g(t,x) = \int_t^{t_{n+1}} \inf_{\pi\in\Pi}\left\{r_\pi(s,x) + Q_\pi g(s,x) - \alpha g(s,x)\right\} ds + V^\alpha_{\text{opt}}(t_{n+1},x), \qquad (t,x) \in \mathcal{T}_n\times\mathfrak{X},$$

によって定義する。ここで V^α_{opt} は最適値関数である。また、可測性に関する次のような仮定を課すことにする。

Assumption B 各 $g \in B(\mathcal{T}_n \times \mathfrak{X})$ に対して、次のように仮定する。

(i) 各 $x \in \mathfrak{X}$ に対して、$\inf_{\pi \in \Pi}\{r_\pi(\cdot, x) + Q_\pi g(\cdot, x)\}$ は $\mathcal{T}_n$ 上の連続関数である。

(ii) 各 $t \in \mathcal{T}_n$ に対して、$\inf_{\pi \in \Pi}\{r_\pi(t, \cdot) + Q_\pi g(t, \cdot)\}$ は $\mathfrak{X}$ 上の可測関数である。

(iii) $S_n g$ は $\mathcal{T}_n \times \mathfrak{X}$ 上の可測関数である。

Lemma 3.4

(i) 各 $n \geq 0$ に対して、前述の作用素 S_n は $B(\mathcal{T}_n \times \mathfrak{X})$ の中に唯一つの不動点 g_n^* をもつ。

(ii) $\sup_{n \geq 0} \|g_n^*\|_n < +\infty$ である。ただし、このノルムの意味は

$$\|g_n^*\| = \sup_{(t,x) \in \mathcal{T}_n \times \mathfrak{X}} |g_n^*(t, x)|$$

Proof. (3.5) を用いることにより、任意の $\pi \in \Pi$、$g \in B(\mathcal{T}_n \times \mathfrak{X})$、$x \in \mathfrak{X}$ に対して、

$$|Q_\pi g(t, x)| \leq 2c(t)\|g\|_n \tag{3.12}$$

が得られる。r_π の有界性と時刻列 $\{t_n\}_{n \geq 0}$ の選び方から、任意の $g \in B(\mathcal{T}_n \times \mathfrak{X})$ に対して

$$\begin{aligned} |S_n g(t, x)| &\leq \int_{t_n}^{t_{n+1}} \left[M + 2c(s)\|g\|_n + \alpha\|g\|_n\right] ds + \frac{M}{\alpha} \\ &\leq (t_{n+1} - t_n)M + \lambda\|g\|_n + \frac{M}{\alpha} < +\infty \end{aligned} \tag{3.13}$$

が成り立ち、$S_n g \in B(\mathcal{T}_n \times \mathfrak{X})$ である。

(3.12) から、任意の $f, g \in B(\mathcal{T}_n \times \mathfrak{X})$、$(t, x) \in \mathcal{T}_n \times \mathfrak{X}$ に対して、

$$\begin{aligned} |S_n f(t, x) - S_n g(t, x)| &\leq \int_t^{t_{n+1}} \Big| \inf_{\pi \in \Pi} \{r_\pi(s, x) + Q_\pi f(s, x) - \alpha f(s, x)\} \\ &\qquad - \inf_{\pi \in \Pi} \{r_\pi(s, x) + Q_\pi g(s, x) - \alpha g(s, x)\} \Big| ds \\ &\leq \int_{t_n}^{t_{n+1}} \sup_{\pi \in \Pi} \Big| Q_\pi(f - g)(s, x) - \alpha\{f(s, x) - g(s, x)\} \Big| ds \\ &\leq \int_{t_n}^{t_{n+1}} \Big\{ \sup_{\pi \in \Pi} |Q_\pi(f - g)(s, x)| + \alpha\|f - g\|_n \Big\} ds \\ &\leq \|f - g\|_n \int_{t_n}^{t_{n+1}} \{\alpha + 2c(s)\} ds \leq \lambda\|f - g\|_n \end{aligned}$$

従って、$\|S_n f - S_n g\|_n \leq \lambda\|f - g\|_n$ が得られ、$0 \leq \lambda < 1/2$ であったので、$S_n : B(\mathcal{T}_n \times \mathfrak{X}) \to B(\mathcal{T}_n \times \mathfrak{X})$ は縮小写像である。縮小写像の不動点定理 (Banach fixed point theorem) により、S_n は $B(\mathcal{T}_n \times \mathfrak{X})$ において唯一つの不動点 g_n^* をもつ。

次に、(3.13) から

$$\|g_n^*\|_n = \|S_n g_n^*\|_n \leq \left(t_{n+1} - t_n + \frac{1}{\alpha}\right) M + \lambda \|g_n^*\|_n$$

となるので、

$$\|g_n^*\|_n \leq \frac{M(\alpha t_{n+1} - \alpha t_n + 1)}{\alpha(1-\lambda)}$$

を得る。ここで、$\int_{t_n}^{t_{n+1}} \{\alpha + 2c(s)\} ds \leq \lambda$、$c(t) \geq 0$ であるから、$\alpha(t_{n+1} - t_n) \leq \lambda$ が成り立つ。これらの不等式から、

$$\sup_{n \geq 0} \|g_n^*\|_n \leq \frac{M(1+\lambda)}{\alpha(1-\lambda)}$$

が得られる。 □

Lemma 3.5 *Lemma 3.4* の不動点 g_n^* は、$(t,x) \in \mathcal{T}_n \times \mathfrak{X}$ に対して最適方程式 *(3.11)* を満たし、更に等式 $g_n^*(t_{n+1}, x) = V_{\mathrm{opt}}^{\alpha}(t_{n+1}, x)$ が成り立つような $\mathcal{D}(\mathcal{T}_n \times \mathfrak{X})$ における唯一つの関数である。

Proof. Lemma 3.4 から作用素 S_n は $B(\mathcal{T}_n \times \mathfrak{X})$ において、唯一つの不動点 g_n^* をもつ。つまり、全ての $(t,x) \in \mathcal{T}_n \times \mathfrak{X}$ に対して、

$$g_n^*(t,x) = \int_t^{t_{n+1}} \inf_{\pi \in \Pi} \left\{ r_\pi(s,x) + Q_\pi g_n^*(s,x) - \alpha g_n^*(s,x) \right\} ds + V_{\mathrm{opt}}^{\alpha}(t_{n+1}, x) \tag{3.14}$$

が成り立つ。(3.14) から関数 $g_n^*(\cdot, x)$ の各 $t \in \mathcal{T}_n$ における微分係数 $D_t g_n^*(t,x)$ が存在し、次の等式が成り立つ。

$$D_t g_n^*(t,x) = -\inf_{\pi \in \Pi} \left\{ r_\pi(t,x) + Q_\pi g_n^*(t,x) - \alpha g_n^*(t,x) \right\} \quad (t,x) \in \mathcal{T}_n \times \mathfrak{X} \tag{3.15}$$

これは、g_n^* が各 $(t,x) \in \mathcal{T}_n \times \mathfrak{X}$ に対して最適方程式 (3.11) を満たすことを示している。また、(3.12), (3.15) から

$$|D_t g_n^*(t,x)| \leq M + \{\alpha + 2c(t)\} \|g_n^*\|_n \tag{3.16}$$

を得る。Assumption B (i) (ii) と等式 (3.14), (3.16) から、g_n^* は集合 $\mathcal{D}(\mathcal{T}_n \times \mathfrak{X})$ に属す。

次に、一意性を示すことにする。関数 $g \in \mathcal{D}(\mathcal{T}_n \times \mathfrak{X})$ が全ての $(t,x) \in \mathcal{T}_n \times \mathfrak{X}$ に対して最適方程式を満たし、さらに $g(t_{n+1}, x) = V_{\mathrm{opt}}^{\alpha}(t_{n+1}, x)$ であると仮定する。このとき、

$$D_t g(s,x) = -\inf_{\pi \in \Pi} \left\{ r_\pi(s,x) + Q_\pi g(s,x) - \alpha g(s,x) \right\}$$

であり、これを $s = t \in \mathcal{T}_n$ から $s = t_{n+1}$ まで積分することにより $\mathcal{T}_n \times \mathfrak{X}$ 上 $g = S_n g$ が成り立つ。Lemma 3.4 により、$\mathcal{T}_n \times \mathfrak{X}$ 上 $g = g_n^*$ である。 □

Theorem 3.2 最適方程式 *(3.11)* を満たす関数が $\mathcal{D}(\mathcal{T}\times\mathfrak{X})$ の中に存在する。

Proof. Lemma 3.5 の g_n^* に対して、関数 $g^*:\mathcal{T}\times\mathfrak{X}\to\mathbb{R}$ を

$$g^*(t,x)=\sum_{n=0}^{\infty} I_{[t_n,t_{n+1})}(t)g_n^*(t,x)$$

によって定義する。ここで $I_{[t_n,t_{n+1})}$ は $\mathcal{T}$ 上の定義関数である。(3.15) から、g^* は最適方程式 (3.11) を満たす。

この関数 g^* が $\mathcal{D}(\mathcal{T}\times\mathfrak{X})$ に属すことを示すことにしよう。まず、Lemma 3.4 の後半より

$$|g^*(t,x)|\le \sup_{n\ge 0}\|g_n^*\|_n<+\infty$$

が成り立つので、$g^*\in B(\mathcal{T}\times\mathfrak{X})$。また、等式 (3.16) から

$$|D_t g^*(t,x)|\le M+\sup_{n\ge 1}\|g_n^*\|_n\{\alpha+2c(t)\}$$

が成り立ち、Assumption B(i)(ii) から、g^* は集合 $\mathcal{D}(\mathcal{T}\times\mathfrak{X})$ の条件 (ii) を満たす。あとは、各 $x\in\mathfrak{X}$ に対して、$g^*(\cdot,x)$ が $\mathcal{T}$ 上の連続関数であることを示せばよい。等式 (3.14) から、関数 $g^*(\cdot,x)$ は $\mathcal{T}$ 上、右側連続であり、かつ $\mathcal{T}\setminus\{t_0,t_1,t_2,\cdots\}$ 上、左側連続である。

そこで、関数 $g^*(\cdot,x)$ が各 t_n において左側連続であることを示すために、

$$\widehat{t}_n=t_n+\inf_{k\ge 0}(t_{k+1}-t_k),\qquad \widehat{\mathcal{T}}_n=[\widehat{t}_n,\widehat{t}_{n+1}],\quad n\ge 0$$

とおく。このとき、

$$\int_{\widehat{t}_n}^{\widehat{t}_{n+1}}\{\alpha+2c(s)\}ds\le 2\lambda<1$$

が成り立つので、Lemma 3.4、3.5 の証明と同様の議論により、$n\ge 1$ に対して、次の 2 つの等式を満たす関数$\widehat{g}_{n-1}^*$ が $\mathcal{D}(\widehat{\mathcal{T}}_{n-1}\times\mathfrak{X})$ において唯一つ存在する。

$\forall x\in\mathfrak{X}$,

$$\begin{cases}\widehat{g}_{n-1}^*(\widehat{t}_n,x)=V_{\mathrm{opt}}^{\alpha}(\widehat{t}_n,x)\\ \alpha\widehat{g}_{n-1}^*(t,x)=\inf_{\pi\in\Pi}\Big\{r_\pi(t,x)+Q_\pi\widehat{g}_{n-1}^*(t,x)+D_t\widehat{g}_{n-1}^*(t,x)\Big\},\quad t\in\widehat{\mathcal{T}}_{n-1}\end{cases}$$

さらに、$\widehat{g}_{n-1}^*$ の一意性と Lemma 3.5から、

$$\widehat{g}_{n-1}^*(t,x)=\begin{cases}g_{n-1}^*(t,x) & \text{if } t\in[\widehat{t}_{n-1},t_n)\\ g_n^*(t,x) & \text{if } t\in[t_n,\widehat{t}_n]\end{cases}$$

が成り立つ。従って、関数 $\widehat{g}_{n-1}^*(\cdot,x)$ の t_n における連続性から、

$$\begin{aligned}\lim_{h\downarrow 0}|g^*(t_n-h,x)-g^*(t_n,x)|&=\lim_{h\downarrow 0}|g_{n-1}^*(t_n-h,x)-g_n^*(t_n,x)|\\&=\lim_{h\downarrow 0}|\widehat{g}_{n-1}^*(t_n-h,x)-\widehat{g}_{n-1}^*(t_n,x)|=0\end{aligned}$$

が得られ、よって関数 $g^*(\cdot,x)$ は各 t_n において左側連続である。以上から、$g^*(\cdot,x)$ は $\mathcal{T}$ 上の連続関数であり、g^* は集合 $\mathcal{D}(\mathcal{T}\times\mathcal{X})$ に属すことが示された。 □

ε-最適ポリシーの存在性を示すために次のような条件を必要とする。

Assumption C 任意の定数 $\varepsilon>0$ と最適方程式 (3.11) の解 $g\in\mathcal{D}(\mathcal{T}\times\mathcal{X})$ に対して、あるポリシー $\pi_\varepsilon\in\Pi$ が存在して全ての $x\in\mathcal{X}$ に対して、

$$\alpha g(t,x)\geq r_{\pi_\varepsilon}(t,x)+Q_{\pi_\varepsilon}g(t,x)+D_tg(t,x)-\varepsilon \qquad \text{a.a.}\, t\in\mathcal{T}$$

が成り立つ。

Theorem 3.3 *Assumption C* を仮定するとき、次が成り立つ。

(i) $V^\alpha_{\rm opt}$ は最適方程式 *(3.11)* の $\mathcal{D}(\mathcal{T}\times\mathcal{X})$ における一意解である。

(ii) 任意の定数 $\varepsilon>0$ に対して、ε-最適ポリシーが存在する。

Proof. $g\in\mathcal{D}(\mathcal{T}\times\mathcal{X})$ を最適方程式 (3.11) の解とする。このとき、全ての $\pi\in\Pi$ と $x\in\mathcal{X}$ に対して、

$$\alpha g(t,x)\leq r_\pi(t,x)+Q_\pi g(t,x)+D_tg(t,x) \qquad \text{a.a.}\, t\in\mathcal{T}$$

が成り立ち、各 $\pi\in\Pi$ に対して、この不等式に Lemma 3.2 を適用することにより、

$$g\leq V^\alpha_{\rm opt} \tag{3.17}$$

が得られる。一方、Assumption Cにより、任意の $\varepsilon>0$ に対して、あるポリシー $\pi_\varepsilon\in\Pi$ が存在して、全ての $x\in\mathcal{X}$ に対して、

$$\alpha g(t,x)\geq r_{\pi_\varepsilon}(t,x)+Q_{\pi_\varepsilon}g(t,x)+D_tg(t,x)-\alpha\varepsilon \qquad \text{a.a.}\, t\in\mathcal{T}$$

が成り立つ。この不等式に Lemma 3.2 を適用することにより、次の不等式が得られる。

$$g\geq V^\alpha_{\pi_\varepsilon}-\varepsilon 1\geq V^\alpha_{\rm opt}-\varepsilon 1 \tag{3.18}$$

ここで $\varepsilon\downarrow 0$ とすると、$g\geq V^\alpha_{\rm opt}$ が得られる。(3.17) から、$g=V^\alpha_{\rm opt}$ となり、この定理の最初の部分が示された。

さらに、(3.18) より各 $\varepsilon>0$ に対して、$V^\alpha_{\pi_\varepsilon}\leq V^\alpha_{\rm opt}+\varepsilon 1$ 成り立つのでポリシー π_ε は ε-最適である。 □

Assumption A、B、Cより弱い条件のもとで、総期待割引損失 V^α_π が (3.10) の一意解であることを示すことができる。各 $n\geq 0$、$\pi\in\Pi$ に対して、

$$S^\pi_nG(t,x)=\int_t^{t_{n+1}}\Big\{r_\pi(s,x)+Q_\pi G(s,x)-\alpha G(s,x)\Big\}ds+V^\alpha_\pi(t_{n+1},x), \qquad (t,x)\in\mathcal{T}_n\times\mathcal{X}$$

によって定義される $B(\mathcal{T}_n\times\mathcal{X})$ 上の作用素 S^π_n を導入する。

Theorem 3.4 各 $\pi \in \Pi$ に対して、

(i) *Assumption A* の *(ii)(b)* において、関数 q_π は、ある連続関数 c_π（ポリシー π に依存してもよい）に対して $q_\pi(t, x\,; \{x\}) \geq -c_\pi(t)$ を満たす。

(ii) 各 $n \geq 0$ に対して、$S_n^\pi G$ は $\mathcal{T}_n \times \mathcal{X}$ 上の可測関数。

を仮定するとき、総期待割引損失 V_π^α は方程式 *(3.10)* の $\mathcal{D}(\mathcal{T} \times \mathcal{X})$ における一意解である。

Proof. Theorem 3.1から、$\mathcal{D}(\mathcal{T} \times \mathcal{X})$ における方程式 (3.10) の任意の解 V は、関数 V_π^α に等しい。また、Theorem 3.2において許容ポリシーの集合 Π を1点 π から成る集合 $\{\pi\}$ であると考えることにより、方程式 (3.10) の解の存在を示すことができる。 □

References

[1] P. Billingsley (1979) *Probability and Measure*, John Wiley & Sons.

[2] B. T. Doshi (1976) *Continuous time control of Markov processes on an arbitrary state space : discounted rewards,* Ann. Statist. 4, 1219-1235.

[3] R. M. Dudley (1989) *Real Analysis and Probability,* Wadsworth & Brooks.

[4] A. Friedman (1975) *Stochastic Differential Equations and Applications Volume 1,* Academic Press.

[5] P. K. Kakumanu (1971) *Continuously discounted Markov decision model with countable state and action spaces,* Ann. Math. Statist. 42, No. 3, 919-926.

[6] H. C. Lai and K. Tanaka (1991) *On continuous-time discounted stochastic dynamic programming,* Appl. Math. Optim. 23, 155-169.

[7] B. L. Miller (1968) *Finite state continuous time Markov decision processes with an infinite planning horizon,* J. Math. Anal. Appl. 22, 552-569.

[8] H. Qiying (1993) *Nonstationary continuous time Markov decision processes with discounted criterion,* J. Math. Anal. Appl. 180, 60-70.