F-041

マルコフ決定過程に基づくマルチエージェントシステムの漸近的性質

An Asymptotic Property of Multiagent Systems Based on a Markov Decision Process

岩田 一貴 † Iwata Kazunori　**池田 和司** ‡ Ikeda Kazushi　**酒井 英昭** ‡ Sakai Hideaki

1 はじめに

複数のエージェントが，互いに協調することにより，ある与えられたタスクに関して最適な政策を学習するようなマルチエージェントシステム (MAS) を考える．MAS が従う確率過程（枠組み）としては，Temporal Difference (TD) 学習などの代表的な強化学習アルゴリズムが適用し易いなどの理由から，マルコフ決定過程（MDP）がよく知られている（例えば，[1] を参照）．本論では，MAS が定常エルゴード MDP に従う場合に成立する漸近的性質を示し，最適政策の学習により MAS の収益が最大化される過程を確率的な観点から明らかにする．

2 準備

初めに記号の準備をする．今，同じ環境に存在する任意の M 人のエージェントに 1 から M の番号をつけ, 番号の集合を $\mathcal{M} \triangleq \{1,\ldots,M\}$ とする．MAS の状態の集合を $\mathcal{S} \triangleq \{s_1,\ldots,s_I\}$，エージェントがとり得る行動の集合を $\mathcal{A} \triangleq \{a_1,\ldots,a_J\}$，実数の集合 $\mathbb{R}$ を適当に離散化した集合を $\mathbb{R}_0 \triangleq \{r_1,\ldots,r_K\} \subset \mathbb{R}$ とする．よって，$|\mathcal{S}| = I$，$|\mathcal{A}| = J$，$|\mathbb{R}_0| = K$ である．記号の簡単のために，MAS 全体（M 人のエージェント）のある状態を $\boldsymbol{s}_{i_1\ldots i_M} = (s_{i_1},\ldots,s_{i_M}) \in \mathcal{S}^M$ と書く．ここで，m 番目の要素 s_{i_m} ($i_m \in \{1,\ldots,I\}$) はエージェント $m \in \mathcal{M}$ の状態を示す．同様に，エージェント全体のある行動を $\boldsymbol{a}_{j_1\ldots j_M} = (a_{j_1},\ldots,a_{j_M}) \in \mathcal{A}^M$，ある報酬を $\boldsymbol{r}_{k_1\ldots k_M} = (r_{k_1},\ldots,r_{k_M}) \in \mathbb{R}_0^M$ と書く．変数 $t \in \mathbb{N}$ は時間ステップを表し，$s_m(t)$，$a_m(t)$，$r_m(t)$ はそれぞれ時間ステップ t におけるエージェント $m \in \mathcal{M}$ の状態，実行した行動，観測した報酬を示す確率変数とする．同様に，時間ステップ t における MAS 全体の状態，行動，報酬の確率変数をそれぞれ $\boldsymbol{s}(t) \triangleq \{s_m(t)|m \in \mathcal{M}\}$，$\boldsymbol{a}(t) \triangleq \{a_m(t)|m \in \mathcal{M}\}$，$\boldsymbol{r}(t) \triangleq \{r_m(t)|m \in \mathcal{M}\}$ と表す．

2.1 MAS の定常エルゴード MDP

定常エルゴード MDP において，ある初期状態分布 $\left\{\Pr(\boldsymbol{s}(1) = \boldsymbol{s}_{i_1\ldots i_M})|\boldsymbol{s}_{i_1\ldots i_M} \in \mathcal{S}^M\right\}$ が与えられたとき，MAS の確率変数 $\boldsymbol{s}(t)$，$\boldsymbol{a}(t)$，$\boldsymbol{r}(t)$ に関する確率分布は，エージェントの政策行列

$$\boldsymbol{\Gamma}^{\pi} \triangleq \left(\mathbf{P}_{(11\ldots1)},\ldots,\mathbf{P}_{(II\ldots I)}\right)^{\mathsf{T}}, \tag{1}$$

および環境の状態遷移行列 $\boldsymbol{\Gamma}^{\mathrm{T}}$

$$\boldsymbol{\Gamma}^{\mathrm{T}} \triangleq \left(\mathbf{P}_{(11\ldots1,11\ldots1)},\ldots,\mathbf{P}_{(II\ldots I,JJ\ldots J)}\right)^{\mathsf{T}}, \tag{2}$$

で定義される．ただし，$\boldsymbol{\Gamma}^{\pi}$ は $I^M \times J^M$ 行列で，各要素を

$$\mathbf{P}_{(i_1\ldots i_M)} \triangleq (p_{i_1\ldots i_M,11\ldots1},\ldots,p_{i_1\ldots i_M,JJ\ldots J})^{\mathsf{T}}, \tag{3}$$

と書く．同様に，$\boldsymbol{\Gamma}^{\mathrm{T}}$ は $I^M J^M \times I^M K^M$ 行列で，各要素を

$$\mathbf{P}_{(i_1\ldots i_M,j_1\ldots j_M)} \triangleq (p_{i_1\ldots i_M,j_1\ldots j_M,11\ldots1,11\ldots1},\ldots,p_{i_1\ldots i_M,j_1\ldots j_M,II\ldots I,KK\ldots K})^{\mathsf{T}}, \tag{4}$$

と記述する．さらに，(3) の各要素の確率は,

$$p_{i_1\ldots i_M,j_1\ldots j_M} \triangleq \Pr(\boldsymbol{a}(t) = \boldsymbol{a}_{j_1\ldots j_M}|\boldsymbol{s}(t) = \boldsymbol{s}_{i_1\ldots i_M}), \tag{5}$$

を示し，(4) の各要素の確率は,

$$p_{i_1\ldots i_M,j_1\ldots j_M,i'_1\ldots i'_M,k_1\ldots k_M} \triangleq \Pr(\boldsymbol{s}(t+1) = \boldsymbol{s}_{i'_1\ldots i'_M},\boldsymbol{r}(t+1) = \boldsymbol{r}_{k_1\ldots k_M}|\boldsymbol{s}(t) = \boldsymbol{s}_{i_1\ldots i_M},\boldsymbol{a}(t) = \boldsymbol{a}_{j_1\ldots j_M}), \tag{6}$$

を示す．本来は各エージェントが学習する場合，政策行列 $\boldsymbol{\Gamma}^{\pi}$ は時変となる．しかし，簡単のために，本論では政策行列 $\boldsymbol{\Gamma}^{\pi}$ が十分にゆっくりと学習により更新されることを仮定し，MAS が定常エルゴード MDP に従うと見なす．このような仮定に基づく解析は，TD 学習などの確率近似法に基づく学習に対してよく行われる（例えば，[2] を参照）．また，環境の状態遷移行列 $\boldsymbol{\Gamma}^{\mathrm{T}}$ の各要素は定数で，各エージェントは $\boldsymbol{\Gamma}^{\mathrm{T}}$ がわからないものとする．

2.2 定常エルゴード MDP の経験分布

任意の n に対する定常エルゴード MDP の経験系列 $\boldsymbol{x} \in (\mathcal{S} \times \mathcal{A} \times \mathbb{R}_0)^{nM}$ を確率変数列

$$\boldsymbol{s}(1),\boldsymbol{a}(1),\boldsymbol{s}(2),\boldsymbol{r}(2),\boldsymbol{a}(2),\ldots,\boldsymbol{s}(n),\boldsymbol{r}(n),\boldsymbol{a}(n),\boldsymbol{r}(n+1),$$

の観測値とする．ある経験系列 $\boldsymbol{x} \in (\mathcal{S} \times \mathcal{A} \times \mathbb{R}_0)^{nM}$ が与えられたとき，任意の状態 $\boldsymbol{s}_{i_1\ldots i_M} \in \mathcal{S}^M$，状態-行動 $(\boldsymbol{s}_{i_1\ldots i_M},\boldsymbol{a}_{j_1\ldots j_M}) \in \mathcal{S}^M \times \mathcal{A}^M$，状態-行動-次状態- 報酬 $(\boldsymbol{s}_{i_1\ldots i_M},\boldsymbol{a}_{j_1\ldots j_M},\boldsymbol{s}_{i'_1\ldots i'_M},\boldsymbol{r}_{k_1\ldots k_M}) \in \mathcal{S}^M \times \mathcal{A}^M \times \mathcal{S}^M \times \mathbb{R}_0^M$ の出現回数をそれぞれ

$$n_{i_1\ldots i_M} \triangleq |\{t \mid \boldsymbol{s}(t) = \boldsymbol{s}_{i_1\ldots i_M},t = 1,\ldots,n\}|, \tag{7}$$

$$n_{i_1\ldots i_M,j_1\ldots j_M} \triangleq |\{t \mid (\boldsymbol{s}(t),\boldsymbol{a}(t)) = (\boldsymbol{s}_{i_1\ldots i_M},\boldsymbol{a}_{j_1\ldots j_M}),t = 1,\ldots,n\}|, \tag{8}$$

† 広島市立大学情報科学部　〒731-3194 広島市安佐南区大塚東 3–4–1　Email: kiwata@im.hiroshima-cu.ac.jp
‡ 京都大学大学院情報学研究科　〒606-8501 京都市左京区吉田本町
Email: {kazushi,hsakai}@i.kyoto-u.ac.jp

$$n_{i_1\ldots i_M,j_1\ldots j_M,i'_1\ldots i'_M,k_1\ldots k_M} \triangleq |\{t \mid (\boldsymbol{s}(t),\boldsymbol{a}(t),\boldsymbol{s}(t+1),\boldsymbol{r}(t+1)) = (\boldsymbol{s}_{i_1\ldots i_M},\boldsymbol{a}_{j_1\ldots j_M},\boldsymbol{s}_{i'_1\ldots i'_M},\boldsymbol{r}_{k_1\ldots k_M}), t=1,\ldots,n\}|, \quad (9)$$

と表す．ただし，(9) では $\boldsymbol{s}(n+1)=\boldsymbol{s}(1)$ と約束する．このとき，MAS の状態 $\boldsymbol{s}_{i_1\ldots i_M}$ に関する経験分布 $f_{i_1\ldots i_M}\in[0,1]$（一般にタイプと呼ぶ [3]）と状態-行動 $(\boldsymbol{s}_{i_1\ldots i_M},\boldsymbol{a}_{j_1\ldots j_M})$ に関する同時タイプ $f_{i_1\ldots i_M,j_1\ldots j_M}\in[0,1]$ は

$$n_{i_1\ldots i_M} = nf_{i_1\ldots i_M}, \quad (10)$$

$$n_{i_1\ldots i_M,j_1\ldots j_M} = nf_{i_1\ldots i_M,j_1\ldots j_M}, \quad (11)$$

により定義される．同様に，状態-行動 $(\boldsymbol{s}_{i_1\ldots i_M},\boldsymbol{a}_{j_1\ldots j_M})$ に関する条件付タイプ $g_{i_1\ldots i_M,j_1\ldots j_M}\in[0,1]$ は，

$$n_{i_1\ldots i_M,j_1\ldots j_M} = n_{i_1\ldots i_M}g_{i_1\ldots i_M,j_1\ldots j_M}, \quad (12)$$

状態-行動-次状態-報酬 $(\boldsymbol{s}_{i_1\ldots i_M},\boldsymbol{a}_{j_1\ldots j_M},\boldsymbol{s}_{i'_1\ldots i'_M},\boldsymbol{r}_{k_1\ldots k_M})$ に関するマルコフタイプ $g_{i_1\ldots i_M,j_1\ldots j_M,i'_1\ldots i'_M,k_1\ldots k_M}\in[0,1]$ は，

$$n_{i_1\ldots i_M,j_1\ldots j_M,i'_1\ldots i'_M,k_1\ldots k_M} = n_{i_1\ldots i_M,j_1\ldots j_M}g_{i_1\ldots i_M,j_1\ldots j_M,i'_1\ldots i'_M,k_1\ldots k_M}, \quad (13)$$

により定義される．簡単のために，(10)–(13) で定義されるタイプを要素に持つ行列（もしくはベクトル）をそれぞれ

$$\boldsymbol{F}_{\mathcal{S}^M} \triangleq \{f_{i_1\ldots i_M}\}, \quad (14)$$

$$\boldsymbol{F}_{\mathcal{S}^M\mathcal{A}^M} \triangleq \{f_{i_1\ldots i_M,j_1\ldots j_M}\}, \quad (15)$$

$$\boldsymbol{\Phi}^\pi \triangleq \{g_{i_1\ldots i_M,j_1\ldots j_M}\}, \quad (16)$$

$$\boldsymbol{\Phi}^{\mathrm{T}} \triangleq \left\{g_{i_1\ldots i_M,j_1\ldots j_M,i'_1\ldots i'_M,k_1\ldots k_M}\right\}, \quad (17)$$

と定義する．さらに，状態 $\boldsymbol{s}_{i_1\ldots i_M}$ および状態-行動 $(\boldsymbol{s}_{i_1\ldots i_M},\boldsymbol{a}_{j_1\ldots j_M})$ に関する定常確率（定常エルゴードならば存在は明らか）を

$$\mathbf{V} \triangleq \lim_{n\to\infty}\boldsymbol{F}_{\mathcal{S}^M}, \qquad \mathbf{W} \triangleq \lim_{n\to\infty}\boldsymbol{F}_{\mathcal{S}^M\mathcal{A}^M}, \quad (18)$$

と定義する．$I^M\times J^M$ 行列 $\boldsymbol{\Phi}^\pi$ および $I^MJ^M\times I^MK^M$ 行列 $\boldsymbol{\Phi}^{\mathrm{T}}$ は，それぞれ政策確率行列 $\boldsymbol{\Gamma}^\pi$ および状態遷移行列 $\boldsymbol{\Gamma}^{\mathrm{T}}$ に関するタイプを要素にもつ行列であることに注意されたい．

3 MAS の定常エルゴード MDP における漸近的性質

強化学習などの報酬を基にした教師無し学習の枠組みにおいて，MAS の最適政策は収益（報酬の重み和）を最大にするような政策として定義される [1, 4]．よって，確率的な観点から見た場合，MAS の収益最大化とは経験系列に関する行列 $\boldsymbol{\Phi}^\pi$ が最適な政策行列（多くの場合，決定論的）に収束することである．

定義 1 (収益最大化) 任意の最適な政策行列 $\boldsymbol{\Gamma}^{\pi^\dagger}$ に対して，MAS の経験系列 $\boldsymbol{x}\in(\mathcal{S}\times\mathcal{A}\times\mathbb{R}_0)^{nM}$ に関する行列 $\boldsymbol{\Phi}^\pi$ が

$$\lim_{n\to\infty}\Pr\left(\boldsymbol{\Phi}^\pi=\boldsymbol{\Gamma}^{\pi^\dagger}\right)=1, \quad (19)$$

を満たすとき，確率収束という意味での収益最大化と呼ぶ．

この定義は [4, Def. 4] に相当する．Sanov の定理を定常エルゴード MDP に当てはめると，n に関する指数的な収束の速さ

$$\lim_{n\to\infty}\frac{1}{n}\log\Pr\left(\boldsymbol{\Phi}^\pi=\boldsymbol{\Gamma}^{\pi^\dagger}\right) = -\mathrm{D}(\boldsymbol{\Gamma}^{\pi^\dagger}\|\boldsymbol{\Gamma}^\pi|\mathbf{V}), \quad (20)$$

が得られ，明らかに $n\to\infty$ に対して $\boldsymbol{\Gamma}^\pi\to\boldsymbol{\Gamma}^{\pi^\dagger}$ となれば収益最大化は達成されることがわかる．ただし，関数 D は

$$\mathrm{D}(\boldsymbol{\Gamma}^{\pi^\dagger}\|\boldsymbol{\Gamma}^\pi|\mathbf{V}) = \sum_{i_1\ldots i_M} v_{i_1\ldots i_M} \times \sum_{i_1\ldots i_M,j_1\ldots j_M} p^\dagger_{i_1\ldots i_M,j_1\ldots j_M}\log\frac{p^\dagger_{i_1\ldots i_M,j_1\ldots j_M}}{p_{i_1\ldots i_M,j_1\ldots j_M}}, \quad (21)$$

を示し，$v_{i_1\ldots i_M}$ は $\mathbf{V}$ の要素，$p^\dagger_{i_1\ldots i_M,j_1\ldots j_M}$ は $\boldsymbol{\Gamma}^{\pi^\dagger}$ の要素を表す．今，別の形で (20) を詳しく考察する．漸近等分割性より，

$$\lim_{n\to\infty}\Pr\left(\boldsymbol{\Phi}^\pi=\boldsymbol{\Gamma}^\pi,\boldsymbol{\Phi}^{\mathrm{T}}=\boldsymbol{\Gamma}^{\mathrm{T}}\right)=1, \quad (22)$$

$$\lim_{n\to\infty}\frac{1}{n}\log P(\boldsymbol{x}) = -\mathrm{H}(\boldsymbol{\Gamma}^\pi|\mathbf{V})-\mathrm{H}(\boldsymbol{\Gamma}^{\mathrm{T}}|\mathbf{W}), \quad (23)$$

$$\lim_{n\to\infty}\frac{1}{n}\log\left|\left\{\boldsymbol{x}\in(\mathcal{S}\times\mathcal{A}\times\mathbb{R}_0)^{nM}\middle|\,\boldsymbol{\Phi}^\pi=\boldsymbol{\Gamma}^\pi,\boldsymbol{\Phi}^{\mathrm{T}}=\boldsymbol{\Gamma}^{\mathrm{T}}\right\}\right| = \mathrm{H}(\boldsymbol{\Gamma}^\pi|\mathbf{V})+\mathrm{H}(\boldsymbol{\Gamma}^{\mathrm{T}}|\mathbf{W}), \quad (24)$$

が成り立つ．ただし，関数 H はエントロピーを表し，例えば (23) における $\mathrm{H}(\boldsymbol{\Gamma}^\pi|\mathbf{V})$ は，$\mathbf{V}$ の要素を $v_{i_1\ldots i_M}$ とすると

$$\mathrm{H}(\boldsymbol{\Gamma}^\pi|\mathbf{V}) = \sum_{i_1\ldots i_M} v_{i_1\ldots i_M} \times \sum_{i_1\ldots i_M,j_1\ldots j_M} p_{i_1\ldots i_M,j_1\ldots j_M}\log\frac{1}{p_{i_1\ldots i_M,j_1\ldots j_M}}, \quad (25)$$

を示す．タイプ理論 [3] により，(22)–(24) は詳細な形で [4, Theo. 1–3] に示されているので証明は割愛する．任意の学習により $\boldsymbol{\Gamma}^\pi\approx\boldsymbol{\Gamma}^{\pi^\dagger}$ となるならば，(22)–(24) より有限だが十分大きな時間ステップ n における収益最大化の確率は，典型集合に含まれる $\mathcal{X}^\dagger_n=\left\{\boldsymbol{x}\in(\mathcal{S}\times\mathcal{A}\times\mathbb{R}_0)^{nM}|\boldsymbol{\Phi}^\pi=\boldsymbol{\Gamma}^{\pi^\dagger}\right\}$ に対して，

$$\Pr(\boldsymbol{\Phi}^\pi=\boldsymbol{\Gamma}^{\pi^\dagger}) \approx \frac{|\,\mathcal{X}^\dagger_n\,|}{\exp\{n\,(\mathrm{H}(\boldsymbol{\Gamma}^\pi|\mathbf{V})+\mathrm{H}(\boldsymbol{\Gamma}^{\mathrm{T}}|\mathbf{W})+\epsilon_n)\}}, \quad (26)$$

となる．ただし，ϵ_n は $\epsilon_n\to0$ $(n\to\infty)$ となる数である．MAS を MDP で定常エルゴードと見なせるほどゆっくりと学習させた場合，例えば最適政策行列 $\boldsymbol{\Gamma}^{\pi^\dagger}$ の各要素が決定論的（0 もしくは 1）ならば，明らかに $\mathrm{H}(\boldsymbol{\Gamma}^\pi|\mathbf{V})\to0$ となる．その結果，(26) において，収益最大化の確率の指数的増加の過程が確認できる．

4 まとめ

MAS が定常エルゴード MDP に従う場合に成立する漸近的性質 (22)–(24) を示し，最適政策の学習により MAS の収益が最大化される過程を (26) により明らかにした．

参考文献

[1] G. Chen *et al.*, Coordinating Multiple Agents via Reinforcement Learning, *Autonomous Agents and Multi-Agent Systems*, vol. 10, no. 3, pp. 273–328, 2005.

[2] H. J. Kushner and G. G. Yin, *Stochastic Approximation Algorithms and Applications*, Springer-Verlag, 1997.

[3] I. Csiszár, The Method of Types, *IEEE Trans. Information Theory*, vol. 44, no. 6, pp. 2505–2523, 1998.

[4] K. Iwata *et al.*, A Statistical Property of Multiagent Learning Based on Markov Decision Process, *IEEE Trans. Neural Networks*, vol. 17, no. 4, pp. 829–842, 2006.