2B-04

二人零和マルコフゲームにおける状態抽象化法に関する研究 *

石橋 宙希 † 島野 雄貴 ‡ 阿部 拳之 § 岩崎 敦 ¶
電気通信大学 無所属 サイバーエージェント 電気通信大学

1 はじめに

本研究は，規模が大きい二人零和マルコフゲームの均衡解を計算するため，マルコフゲームの状態を抽象化する方法について吟味する．二人零和マルコフゲームとは，エージェントの利得がお互いの行動だけでなく，環境を表す状態によって決まるゲームであり，その状態遷移はマルコフ過程に従う．例えば，サッカーやアメリカンフットボールのようなゲームでは，場面場面の状態によって行動の価値が変わるため，マルコフゲームとして記述するのが望ましい．しかし，そのようなゲームの状態数はゲームの要素の数に対して指数的に増加するため，その均衡計算が困難になる．そこで本研究では，シングルエージェントであるマルコフ決定過程 (Markov Decision Process, MDP) の状態を抽象化する方法 [2] を拡張し，どの程度の情報が失われるかを吟味する．

2 二人零和マルコフゲームの状態抽象化

二人零和マルコフゲームとは，二人のプレイヤが受け取る報酬の和が 0 であるようなマルコフゲームを指す [3, 1]．二人零和マルコフゲームを $E := (N, S, A^1, A^2, R, P, \gamma, \rho_0)$ と定義する．二人のプレイヤを $i \in N = \{1, 2\}$ として，期間 $t = 0, 1, 2, \cdots, T$ に渡って二人零和マルコフゲームを繰り返す．プレイヤ 1, 2 は状態

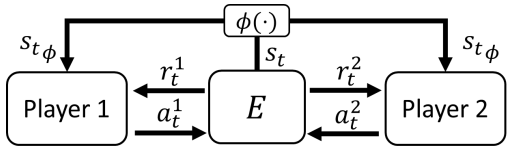

図 1: 状態抽象化した二人零和マルコフゲーム

$s_t \in S$ を受け取り，それらをもとに行動 $a_t^1 \in A^1, a_t^2 \in A^2$ を選択する．そして，各プレイヤは行動の組 $\boldsymbol{a}_t = (a_t^1, a_t^2)$ をもとに報酬関数 $R : S \times A^1 \times A^2 \to [R_{\min}, R_{\max}]$ に従って報酬 r_t を得て，遷移関数 $P : S \times A^1 \times A^2 \to \Delta(S)$ に従って次の状態 s_{t+1} に遷移する．初期状態分布は $\rho_0 \in \Delta(S)$ とする．また，ある状態でのプレイヤ i の振る舞いを示すものを方策 $\pi_i(a^i|s) \in \Pi_i : S \to \Delta(A^i)$ と定義して，二人のプレイヤの方策の組を $\boldsymbol{\pi} = (\pi_1, \pi_2)$ とする．このような二人零和マルコフゲームにおいて，各プレイヤの目標は報酬 r_t と割引因子 $\gamma \in [0, 1)$ によって決定される割引利得和の最大化である．

各プレイヤが状態 s_t において行動 a_t^1, a_t^2 を選択し，その後は方策 π_1, π_2 に従う場合の行動価値関数 Q を，

$$Q_1^{\boldsymbol{\pi}}(s_t, \boldsymbol{a}_t) = R(s_t, \boldsymbol{a}_t) + \gamma \sum_{s_{t+1} \in S} P(s_{t+1}|s_t, \boldsymbol{a}_t) Q_1^{\boldsymbol{\pi}}(s_{t+1}, \boldsymbol{\pi}(s_{t+1})),$$
$$Q_2^{\boldsymbol{\pi}}(s_t, \boldsymbol{a}_t) = -Q_1^{\boldsymbol{\pi}}(s_t, \boldsymbol{a}_t),$$

と定義して，各プレイヤが状態 s_t において方策 π_1, π_2 に従って行動を選択する場合の価値関数 V を，

$$V_1^{\boldsymbol{\pi}}(s_t) = Q_1^{\boldsymbol{\pi}}(s_t, \boldsymbol{\pi}(s_t)), V_2^{\boldsymbol{\pi}}(s_t) = -V_1^{\boldsymbol{\pi}}(s_t),$$

と定義する．また，二人零和マルコフゲームにおける均衡方策 $\boldsymbol{\pi}^* = (\pi_1^*, \pi_2^*)$ を，

$$\forall \pi_1, \pi_2 : V_1^{\pi_1^*, \pi_2}(s) \geq V_1^{\pi_1^*, \pi_2^*}(s) \geq V_1^{\pi_1, \pi_2^*}(s),$$

と定義する [1]．本研究ではこの均衡方策を求めるアルゴリズムとして，ミニマックス Q 学習を用いる [3]．

二人零和マルコフゲームの具体例として，二人のプレイヤが 4×5 のフィールド上で行うマルコフサッカーを考える [1, 3]．プレイヤの初期位置は図 2 のとおりであり，ボールの所持者はランダムとする．各期において，すべてのプレイヤは同時に上，右，下，左への 1 マス移動と停止の 5 つから行動を選択する．プレイヤがボールを所持したままゴールのマスに移動したとき，そのプレイヤは報酬 1 を得て，相手のプレイヤは報酬 −1 を得る．

---

* Solving Two-Player Zero-Sum Markov Game via Approximate State Abstraction
† Hiroki Ishibashi, The University of Electro-Communications
‡ Yuki Shimano, Freelance
§ Kenshi Abe, CyberAgent, Inc.
¶ Atsushi Iwasaki, The University of Electro-Communications

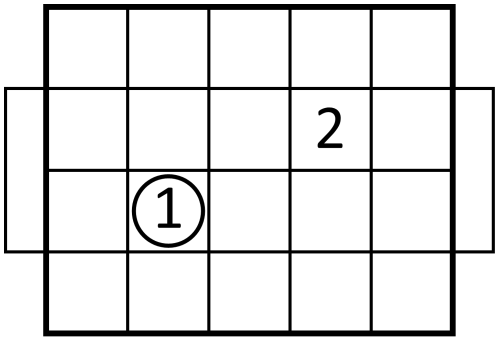

図 2: マルコフサッカーの初期状態

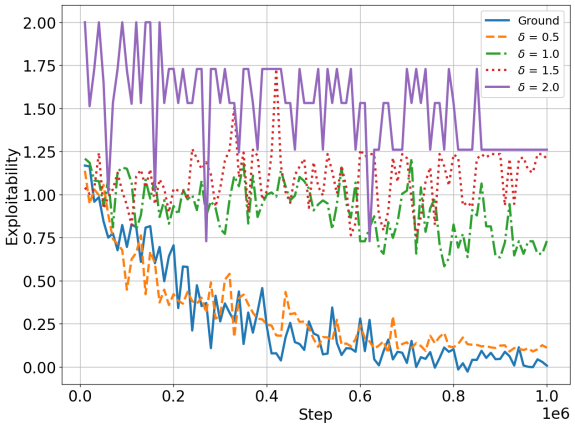

図 3: 状態抽象化と Exploitability

文献 [2] で紹介されている， MDP において行動価値関数 Q をもとに状態を抽象化する手法を二人零和マルコフゲームに拡張する．状態抽象化関数を $\phi : S \to S_\phi$ とし，任意の状態 s を抽象化した状態を $s_\phi = \phi(s)$ と定義する．また，状態抽象化された二人零和マルコフゲームを $E_\phi := \langle N, S_\phi, A^1, A^2, R_\phi, P_\phi, \gamma, \rho_0^\phi \rangle$ とする．任意の状態 s^1, s^2 に対して，任意の 0 以上の実数 δ を用いて状態抽象化関数 $\phi_{Q_\epsilon^{\pi^*}}$ を，

$$\begin{aligned} &\phi_{Q_\delta^{\pi^*}}(s^1) = \phi_{Q_\delta^{\pi^*}}(s^2) \\ &\Rightarrow \forall i, \boldsymbol{a} : |Q_i^{\pi^*}(s^1, \boldsymbol{a}) - Q_i^{\pi^*}(s^2, \boldsymbol{a})| \leq \delta, \end{aligned} \tag{1}$$

と定義する．ある状態 s において，元のマルコフゲームでの均衡方策 $\boldsymbol{\pi}^*(s)$ と状態抽象化したマルコフゲームでの均衡方策 $\boldsymbol{\pi}^\phi(s_\phi)$ がどれだけ乖離しているかを表す Expolitability を $V^{\exp}(s) = \sum_{i \in N} \max_{\pi_i} V_i^{\pi_i, \pi_{-i}^\phi}(s)$ とする [1]．Exploitability が 0 ならば，$\boldsymbol{\pi}^\phi$ は $\boldsymbol{\pi}^*$ と等しい．

3 計算機実験

マルコフサッカーについて行動価値関数 Q をもとにした状態抽象化の実験を行った．学習のステップ数は 1000000 として，割引因子 γ は 0.9 とした．

元のマルコフサッカーの均衡方策 $\boldsymbol{\pi}^*$ をミニマックス Q 学習を用いて求めた．式 (1) で定義した状態抽象化関数 $\phi_{Q_\delta^{\pi^*}}$ に従って δ が 0.5, 1.0, 1.5, 2.0 の場合について状態抽象化を行い，それぞれの場合の状態数 $|S_{\phi_{Q_\delta^{\pi^*}}}|$ を計算し，ミニマックス Q 学習を用いて均衡方策の組 $\boldsymbol{\pi}^{\phi_{Q_\delta^{\pi^*}}}$ を得た．また，プレイヤ 1 が方策 $\pi_1^{\phi_{Q_\delta^{\pi^*}}}$，プレイヤ 2 が方策 π_2^* に従うマルコフサッカーを行い，各 δ ごとのプレイヤ 1 の勝率を求めた．そして，各 δ ごとの Exploitability の推移を確認した．ここで，真の Exploitability の計算は困難なので Q 学習を適用することで近似的に求めた．

元のマルコフサッカーの状態数 $|S| = 760$ に対し，δ を 0.5, 1.0, 1.5, 2.0 と徐々に増加させたとき，マルコフサッカーの状態数 $|S_{\phi_{Q_\delta^{\pi^*}}}|$ はそれぞれ 576, 304, 94, 1 と削減できた．二人のプレイヤがともに均衡方策 $\boldsymbol{\pi}^*$ に従うとき，プレイヤ 1 の勝率は 48.5% であった．また，プレイヤ 1 が方策 $\pi_1^{\phi_{Q_\delta^{\pi^*}}}$，プレイヤ 2 が方策 π_2^* に従った場合のプレイヤ 1 の勝率は δ が 0.5, 1.0 ,1.5, 2.0 のそれぞれについて 48.8%, 28.8%, 20.3%, 16.7% となった．最後に，図 3 に，状態を抽象化したときの学習の推移を表した．ここで横軸を Step，縦軸を Exploitability とし，状態抽象化していない場合の結果を”Ground” と表した．

4 おわりに

本研究では，規模の大きい二人零和マルコフゲームの均衡解を計算するために，MDP の状態抽象化の手法を二人零和マルコフゲームへ拡張した．プレイヤ 1 の勝率や Exploitability の収束の様子から，δ が 0.5 のときは元のマルコフサッカーに近いと分かった．今後は，より大きな規模のマルコフゲームに今回の状態抽象化を適用し，そのときの性能を評価したい．

参考文献

[1] K. Abe and Y. Kaneko. Off-policy exploitability-evaluation in two-player zero-sum markov games. In *AAMAS*, 2021.

[2] D. Abel, D. Hershkowitz, and M. Littman. Near optimal behavior via approximate state abstraction. In *ICML*, pp. 2915–2923. PMLR, 2016.

[3] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pp. 157–163. Elsevier, 1994.