CF-016

二人零和マルコフゲームにおける状態抽象化に関する研究

石橋 宙希 * 阿部 拳之 † 岩崎 敦 *
Hiroki Ishibashi Kenshi Abe Atsushi Iwasaki

1 はじめに

本研究では，二人零和マルコフゲームに状態抽象化を導入し，その性能を確認する．二人零和マルコフゲームとは，二人のプレイヤの利得が環境を表す状態とお互いの行動によって決まるゲームであり，その状態遷移はマルコフ決定過程に従う．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．二人零和マルコフゲームでは，ゲームの状態数が増加するにつれて均衡計算が困難になる．そこで，二人零和マルコフゲームの状態を抽象化し，その性能を吟味する．状態抽象化とは，複数の異なる状態を 1 つの状態とみなすことで状態数を削減する方法であり，マルコフ決定過程の最適方策を求めたい状況については，多くの先行研究が存在する [3, 4, 5, 6]．本研究は，文献 [2] のマルコフ決定過程の状態抽象化をマルコフゲームに拡張した．このとき，状態抽象化が適用されたゲームでは元のゲームの均衡解と異なる均衡解しか得られないことがある．そこで状態抽象化が適用されたゲームの均衡解を評価するために，元のゲームの均衡解との距離を表す Exploitability [1, 9] の大きさに関するバウンドを導いた．最後に，二人零和マルコフゲームの一つであるマルコフサッカー [7] に状態抽象化を適用して均衡計算を行い，その結果を吟味する．

2 モデル

2.1 二人零和マルコフゲーム

二人零和マルコフゲーム $\mathcal{M}$ を $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}_1, \mathcal{A}_2, P, R, \gamma \rangle$ と定義する．ここで，$\mathcal{S}$ は有限状態集合，$\mathcal{A}_i$ は各プレイヤ $i \in \{1,2\}$ の有限行動集合，$P : \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \to \Delta(\mathcal{S})$ は遷移関数，$R : \mathcal{S} \times \mathcal{A}_1 \times \mathcal{A}_2 \to [-1, 1]$ は報酬関数，$\gamma \in [0, 1)$ は割引因子を示している．以降，プレイヤ i ではないプレイヤを $-i$ と表す．二人のプレイヤの行動の組を $\boldsymbol{a} = (a_i, a_{-i})$ とし，行動の組の集合を $\mathcal{A} = \mathcal{A}_i \times \mathcal{A}_{-i}$ とする．状態 $s \in \mathcal{S}$ で行動 $\boldsymbol{a} \in \mathcal{A}$ が実行されたとき，プレイヤ 1 が得る報酬 R_1 は $R_1(s, \boldsymbol{a}) := R(s, \boldsymbol{a})$ ，プレイヤ 2 が得る報酬 R_2 は $R_2(s, \boldsymbol{a}) := -R(s, \boldsymbol{a})$ とする．

ある状態におけるプレイヤ i の行動の振る舞いを示す方策を $\pi_i : \mathcal{S} \to \Delta(A_i)$ と定義し，二人のプレイヤの方策の組を $\boldsymbol{\pi} = (\pi_i, \pi_{-i})$ とする．ある状態 $s \in \mathcal{S}$ から各プレイヤが特定の方策 $\boldsymbol{\pi}$ に従って行動を選択し続けるときの割引期待利得和を価値関数とし，プレイヤ $i \in \{1,2\}$ の価値関数を，

$$V_i^{\boldsymbol{\pi}}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_i(s_t, \boldsymbol{a}_t) \,\middle|\, s_0 = s, \boldsymbol{a}_t \sim \boldsymbol{\pi}(\cdot|s_t), s_{t+1} \sim P(\cdot|s_t, \boldsymbol{a}_t), \forall t \geq 0\right],$$

と定義する．二人零和マルコフゲームでは，任意の状態 $s \in \mathcal{S}$ について $V_i^{\boldsymbol{\pi}}(s) = -V_{-i}^{\boldsymbol{\pi}}(s)$ が成り立つ．次に，ある状態 $s \in \mathcal{S}$ で各プレイヤが行動 $a_1 \in \mathcal{A}_1, a_2 \in \mathcal{A}_2$ を選択し，その後は特定の方策 $\boldsymbol{\pi}$ に従って行動を選択する場合の割引期待利得和を行動価値関数とし，各プレイヤ $i \in \{1,2\}$ の行動価値関数を，

$$Q_i^{\boldsymbol{\pi}}(s, \boldsymbol{a}) = R_i(s, \boldsymbol{a}) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s, \boldsymbol{a}) V_i^{\boldsymbol{\pi}}(s'),$$

と定義する．価値関数と同様に，任意の状態 $s \in \mathcal{S}$ と行動 $\boldsymbol{a} \in \mathcal{A}$ について $Q_i^{\boldsymbol{\pi}}(s, \boldsymbol{a}) = -Q_{-i}^{\boldsymbol{\pi}}(s, \boldsymbol{a})$ が成り立つ．また，行動価値関数 $Q_i^{\boldsymbol{\pi}}$ を用いて価値関数 $V_i^{\boldsymbol{\pi}}$ を表すことができる：

$$V_i^{\boldsymbol{\pi}}(s) = \sum_{\boldsymbol{a} \in \mathcal{A}} \boldsymbol{\pi}(\boldsymbol{a}|s) Q_i^{\boldsymbol{\pi}}(s, \boldsymbol{a}).$$

2.2 均衡方策

二人零和マルコフゲーム $\mathcal{M}$ において，以下の性質を満たす方策 $\boldsymbol{\pi}^*$ をナッシュ均衡の方策と呼ぶ：

$$\forall s \in \mathcal{S}, \forall \boldsymbol{\pi},\ V_1^{\pi_1^*, \pi_2}(s) \geq V_1^{\boldsymbol{\pi}^*}(s) \geq V_1^{\pi_1, \pi_2^*}(s).$$

二人零和マルコフゲームにおけるナッシュ均衡 $\boldsymbol{\pi}^*$ の状態価値関数 $V_i^{\boldsymbol{\pi}^*}(s)$ は，任意の状態 $s \in \mathcal{S}$ に対して以下の等式を満たす [8]：

$$V_i^{\boldsymbol{\pi}^*}(s) = \max_{p \in \Delta(\mathcal{A}_i)} \min_{a_{-i} \in \mathcal{A}_{-i}} \sum_{a_i \in \mathcal{A}_i} p(a_i) Q_i^{\boldsymbol{\pi}^*}(s, \boldsymbol{a}). \quad (1)$$

また，任意の方策の組 $\boldsymbol{\pi} = (\pi_i, \pi_{-i})$ を評価する指標として，任意の状態 $s \in \mathcal{S}$ に対する Exploitability を以下のように定義する [1, 9]：

$$\text{exploit}(\boldsymbol{\pi}, s) = \sum_{i \in \{1,2\}} \left(\max_{\pi_i'} V_i^{\pi_i', \pi_{-i}}(s) - V_i^{\boldsymbol{\pi}}(s) \right).$$

この Exploitability は，プレイヤが評価したい方策 $\boldsymbol{\pi}$ に従う場合に比べ，方策を変更することでどのくらい価値関数を改善

* 電気通信大学, The University of Electro-Communications
† サイバーエージェント, CyberAgent

することができるかを示す指標であり，$\pi_i^\dagger$ は方策 π_{-i} に対するプレイヤ i の最適方策を示している．定義より，評価したい方策 $\boldsymbol{\pi}$ が均衡方策 $\boldsymbol{\pi}^*$ に近いほど，Exploitability は 0 に近い値をとる．

2.3 ミニマックス Q 学習

本研究では，均衡方策 $\boldsymbol{\pi}^*$ を求めるアルゴリズムにミニマックス Q 学習 [7] を用いた（アルゴリズム 1）．ミニマックス Q 学習とは，ある状態に対する行動価値関数 $Q_i^{\boldsymbol{\pi}}(s, \boldsymbol{a})$ を更新し，それを用いて方策 $\boldsymbol{\pi}(s)$ を更新することで，最終的にすべての状態についての方策が均衡方策に収束するアルゴリズムである．

まず，第 t ステップ目において，ある状態 s_l で各プレイヤが確率 β でランダムに行動を選択し，確率 $1-\beta$ でその時点での方策 π_i に従って行動を選択する．その行動 $a_{1,t}, a_{2,t}$ をもとに，各プレイヤは報酬 $r_{1,t}, r_{2,t}$ を得て，次のステップの状態 s_{l+1} が決まる．次に，状態 s，選択された行動の組 $\boldsymbol{a}$，割引因子 γ，学習率 α_t をもとに各プレイヤの行動価値関数 $Q_i^{\boldsymbol{\pi}}$ を以下の式で更新する．学習率 α_t は行動価値関数 $Q_i^{\boldsymbol{\pi}}$ を更新するときにその時点での行動価値関数 $Q_i^{\boldsymbol{\pi}}$ をどの程度残すかという割合であり，ステップ t に依存する．

$$Q_i^{\boldsymbol{\pi}}(s_t, \boldsymbol{a}_t) := (1-\alpha_t) Q_i^{\boldsymbol{\pi}}(s_t, \boldsymbol{a}_t) + \alpha_t (r_{i,t} + \gamma V_i^{\boldsymbol{\pi}}(s_{t+1})).$$

そして，更新した行動価値関数 $Q_i^{\boldsymbol{\pi}}$ と線形計画法を用いて各プレイヤの方策 π_i を更新し，その方策を用いて各プレイヤの価値関数 $V_i^{\boldsymbol{\pi}}$ を更新する：

$$\pi_i(\cdot|s_t) := \arg\max_{p \in \Delta(\mathcal{A}_i)} \min_{a_{-i,t} \in \mathcal{A}_{-i}} \sum_{a_{i,t} \in A^i} p(a_{i,t}) Q_i^{\boldsymbol{\pi}}(s_t, \boldsymbol{a}_t).$$

$$V_i^{\boldsymbol{\pi}}(s_t) := \min_{a_{-i,t} \in \mathcal{A}_{-i}} \sum_{a_{i,t} \in \mathcal{A}_i} \pi_i(a_{i,t}|s_t) Q_i^{\boldsymbol{\pi}}(s_t, \boldsymbol{a}_t).$$

以上の更新を繰り返すことで，各プレイヤの方策 π_i を均衡方策に近づけることができる．

アルゴリズム 1 ミニマックス Q 学習

1. 初期状態 $s_0 \in \mathcal{S}$ を与える
   $\forall i \in \{1,2\}, i \in \{1,2\}, s \in \mathcal{S}, \boldsymbol{a} \in \mathcal{A} : Q_i^{\boldsymbol{\pi}}(s, \boldsymbol{a}) := 1.0$
   $\forall s \in \mathcal{S} : V_i^{\boldsymbol{\pi}}(s) := 1.0$
   $\forall i \in \{1,2\}, s \in \mathcal{S}, a_i \in \mathcal{A}_i : \pi_i(a_i|s) := |A^i|^{-1}$
2. 以下の操作を $t = 0, 1, \cdots, T$ 回まで繰り返す
   (a) 各プレイヤは，確率 β でランダムに，確率 $1-\beta$ で $\boldsymbol{\pi}(s_t)$ に従って行動 $a_{1,t}, a_{2,t}$ を選択する
   (b) 状態 s_t と行動の組 $\boldsymbol{a}_t$ をもとに，各プレイヤは報酬 $r_{1,t}, r_{2,t}$ を得て，状態 s_{t+1} に遷移する
   (c) 行動価値関数 $Q_i^{\boldsymbol{\pi}}$ を更新する
   $Q_i^{\boldsymbol{\pi}}(s_t, \boldsymbol{a}_t)$
   $\leftarrow (1-\alpha_t) Q_i^{\boldsymbol{\pi}}(s_t, \boldsymbol{a}_t) + \alpha_t (r_{i,t} + \gamma V_i^{\boldsymbol{\pi}}(s_{t+1}))$
   (d) 方策 $\boldsymbol{\pi}$ を更新する
   $\pi_i(\cdot|s_t)$
   $\leftarrow \arg\max_{p \in \Delta(\mathcal{A}_i)} \min_{a_{-i,t}} \sum_{a_{i,t}} p(a_{i,t}) Q_i^{\boldsymbol{\pi}}(s_t, \boldsymbol{a}_t)$
   (e) 価値関数 $V_i^{\boldsymbol{\pi}}$ を更新する
   $V_i^{\boldsymbol{\pi}}(s_t) \leftarrow \min_{a_{-i,t}} \sum_{a_{i,t}} \pi_i(a_{i,t}|s_t) Q_i^{\boldsymbol{\pi}}(s_t, \boldsymbol{a}_t)$
3. 方策 π_1, π_2 を出力する

3 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では，状態数や行動の選択肢が多くなるにつれて均衡計算が困難になるという課題がある．状態抽象化と呼ばれる方法では，複数の異なる状態を 1 つの同じ状態だとしてみなした新しいゲームを考え，そのゲームの均衡方策を代わりに計算することで計算量の削減を行なう．先行研究 [2] では，様々な指標を用いて MDP の状態を抽象化した場合について，その抽象化された MDP での最適な方策が，元の MDP においてどの程度の性能を達成するかについて理論的な分析を行っている．本論文では，この議論を二人零和マルコフゲームへと拡張する．

本研究では，元のゲームの状態集合から状態抽象化が適用されたゲームの状態集合への写像である状態抽象化関数 $\phi : \mathcal{S} \to \mathcal{S}_A$ を用いて状態を抽象化することを考える．なお，$\mathcal{S}_A$ は状態抽象化が適用されたゲームの状態集合である．この状態抽象化関数 $\phi : \mathcal{S} \to \mathcal{S}_A$ のもとで，与えられた状態 s と同じ状態へと抽象化される状態の集合 G を以下のように定義する：

$$G(s) = \begin{cases} \{g \in \mathcal{S} \mid \phi(g) = \phi(s)\} & \text{if } s \in \mathcal{S} \\ \{g \in \mathcal{S} \mid \phi(g) = s\} & \text{if } s \in \mathcal{S}_A \end{cases}.$$

状態抽象化が適用されたゲームの遷移関数と報酬関数を定義するために，重み関数 $w : \mathcal{S} \to [0,1]$ を次のように定義する：

$$\forall s \in \mathcal{S} : \sum_{g \in G(s)} w(g) = 1.$$

この重み関数の具体的な例としては，$w(s) = 1/|G(s)|$ が挙げられる．重み関数を用いて，状態抽象化が適用されたゲームの遷移関数 $P_A : \mathcal{S}_A \times \mathcal{A} \to \Delta(\mathcal{S}_A)$ と報酬関数 $R_A : \mathcal{S}_A \times \mathcal{A} \to \mathbb{R}$ を，

$$P_A(s'_A|s_A, \boldsymbol{a}) = \sum_{s \in G(s_A)} \sum_{s' \in G(s'_A)} P(s'|s, \boldsymbol{a}) w(s),$$

$$R_A(s_A, \boldsymbol{a}) = \sum_{s \in G(s_A)} R(s, \boldsymbol{a}) w(s),$$

と定義する．これらを用いて，状態抽象化が適用された二人零和マルコフゲーム $\mathcal{M}_A$ を $\mathcal{M}_A = \langle \mathcal{S}_A, \mathcal{A}_1, \mathcal{A}_2, P_A, R_A, \gamma \rangle$ と表す．

状態抽象化が適用された二人零和マルコフゲーム $\mathcal{M}_A$ におけるプレイヤ i の方策を $\boldsymbol{\pi}_{A,i} : \mathcal{S}_A \to \Delta(\mathcal{A}_i)$ とし，二人のプレイヤの方策の組を $\boldsymbol{\pi}_A = (\pi_{A,i}, \pi_{A,-i})$ とする．状態抽象化が適用されたゲームにおいて，ある状態 $s_A \in \mathcal{S}_A$ から各プレイヤが方策 $\boldsymbol{\pi}_A \in \Delta(\mathcal{A})$ に従って行動を選択し続ける場合

のプレイヤ i の価値関数を $V_{A,i}^{\boldsymbol{\pi}_A}(s_A)$ とする．同様に，状態抽象化が適用されたゲームにおいて，ある状態 $s_A \in \mathcal{S}_A$ で行動 $\boldsymbol{a} \in \mathcal{A}$ を実行し，その後は方策 $\boldsymbol{\pi}_A \in \Delta(\mathcal{A})$ に従って行動を選択する場合のプレイヤ i の行動価値関数を $Q_{A,i}^{\boldsymbol{\pi}_A}(s_A, \boldsymbol{a})$ とする．$V_{A,i}^{\boldsymbol{\pi}_A}$ と $Q_{A,i}^{\boldsymbol{\pi}_A}$ の定義は，元の二人零和マルコフゲームの価値関数 $V_i^{\boldsymbol{\pi}}$ と行動価値関数 $Q_i^{\boldsymbol{\pi}}$ の遷移関数 P と報酬関数 R をそれぞれ P_A と R_A に置き換えたものである．状態抽象化が適用された二人零和マルコフゲーム $\mathcal{M}_A$ におけるナッシュ均衡 $\boldsymbol{\pi}_A^*$ は以下の不等式を満たす：

$$\forall s_A \in \mathcal{S}_A, \forall \boldsymbol{\pi}_A : \\ V_{A,1}^{\pi_{A,1}^*, \pi_{A,2}}(s_A) \geq V_{A,1}^{\boldsymbol{\pi}_A^*}(s_A) \geq V_{A,1}^{\pi_{A,1}, \pi_{A,2}^*}(s_A).$$

また，状態抽象化が適用されたゲームにおける方策 $\boldsymbol{\pi}_A(\phi(s))$ を元のゲームに適用する際の方策を $\boldsymbol{\pi}_{GA}(s)$ とし，それらの関係を，

$$\forall s \in \mathcal{S} : \boldsymbol{\pi}_{GA}(s) = \boldsymbol{\pi}_A(\phi(s)),$$

とする．元のゲームにおいて，方策 $\boldsymbol{\pi}_{GA}$ に従って行動を選択する場合の価値関数や行動価値関数を $V_i^{\boldsymbol{\pi}_{GA}}, Q_i^{\boldsymbol{\pi}_{GA}}$ とする．

3.1 行動価値関数に基づく状態抽象化

本章では，行動価値関数 $Q_i^{\boldsymbol{\pi}^*}$ に基づいて状態抽象化が行われている場合について検討する．具体的には，状態抽象化関数 ϕ によってある 2 つの状態 $s_1, s_2 \in \mathcal{S}$ が同じ状態に抽象化されるとき，それらの状態におけるミニマックス値 $Q_i^{\boldsymbol{\pi}^*}(s_1, a_1, a_2)$, $Q_i^{\boldsymbol{\pi}^*}(s_2, a_1, a_2)$ が，最大でも ϵ しか離れていない状況を考える：

$$\phi(s_1) = \phi(s_2) \\ \Rightarrow \forall i \in \{1,2\}, \forall \boldsymbol{a} \in \mathcal{A} : \left| Q_i^{\boldsymbol{\pi}^*}(s_1, \boldsymbol{a}) - Q_i^{\boldsymbol{\pi}^*}(s_2, \boldsymbol{a}) \right| \leq \epsilon. \tag{2}$$

ここで，ϵ は任意の非負の実数とする．

以下の定理 3.1では，この仮定を満たすような状態抽象化が適用されたゲームにおける均衡方策 $\boldsymbol{\pi}_{GA}^*$ が，元のゲームにおける均衡方策から ϵ 程度しか離れないことを示している：

定理 3.1. 状態抽象化関数 ψ が式 (2) を満たすとする．このとき，この状態抽象化関数を適用した二人零和マルコフゲームにおける均衡方策 $\boldsymbol{\pi}_{GA}^*$ の，元のゲームでの Exploitability は以下の式で抑えられる：

$$\forall s \in \mathcal{S},\ \text{exploit}\left(\boldsymbol{\pi}_{GA}^*, s\right) \leq \frac{12\epsilon}{(1-\gamma)^3}.$$

3.2 定理 3.1の証明

本章では，定理 3.1の証明を与える．なお，本章で導入する補題の証明は付録に示した．まず，Exploitability の定義から，

$$\text{exploit}\left(\boldsymbol{\pi}_{GA}^*, s\right) \\ = \sum_{i \in \{1,2\}} \left(\max_{\pi_i} V_i^{\pi_i, \pi_{GA,-i}^*}(s) - V_i^{\boldsymbol{\pi}_{GA}^*}(s) \right). \tag{3}$$

したがって，以降では元のゲームにおいて各プレイヤ i が方策 $\pi_{GA,i}^*$ から逸脱するインセンティブ $\max_{\pi_i} V_i^{\pi_i, \pi_{GA,-i}^*}(s) - V_i^{\boldsymbol{\pi}_{GA}^*}(s)$ の上界を導出する．ここで，元のゲームにおける $\pi_{GA,-i}^*$ に対する最適方策を $\pi_i^\dagger$ を定義すると，$\max_{\pi_i} V_i^{\pi_i, \pi_{GA,-i}^*}(s) - V_i^{\boldsymbol{\pi}_{GA}^*}(s)$ に関して以下の補題を得る．

補題 3.2. ある定数 $\delta \geq 0$ が存在し，任意の状態 $s \in \mathcal{S}$ と行動の組 $\boldsymbol{a} \in \mathcal{A}$ について，次の式が成り立つとする：

$$\left| Q_{A,i}^{\boldsymbol{\pi}_A^*}(\phi(s), \boldsymbol{a}) - Q_i^{\pi_i^\dagger, \pi_{GA,-i}^*}(s, \boldsymbol{a}) \right| \leq \delta.$$

この時，任意の状態 $s \in \mathcal{S}$ について次の不等式が成立する：

$$V_i^{\pi_i^\dagger, \pi_{GA,-i}^*}(s) - V_i^{\boldsymbol{\pi}_{GA}^*}(s) \leq \frac{2\delta}{1-\gamma}.$$

この補題から，$\pi_{A,-i}^*$ に対する抽象化したゲームと元のゲームそれぞれにおける最適方策の状態行動価値関数の差 $Q_{A,i}^{\boldsymbol{\pi}_A^*}(\phi(s), \boldsymbol{a}) - Q_i^{\pi_i^\dagger, \pi_{GA,-i}^*}(s, \boldsymbol{a})$ の上・下界の導出が可能ならば，定理の主張を証明できることが分かる．三角不等式から，この差は以下のように分解できる：

$$\left| Q_{A,i}^{\boldsymbol{\pi}_A^*}(\phi(s), \boldsymbol{a}) - Q_i^{\pi_i^\dagger, \pi_{GA,-i}^*}(s, \boldsymbol{a}) \right| \\ \leq \left| Q_{A,i}^{\boldsymbol{\pi}_A^*}(\phi(s), \boldsymbol{a}) - Q_i^{\boldsymbol{\pi}^*}(s, \boldsymbol{a}) \right| \\ + \left| Q_i^{\boldsymbol{\pi}^*}(s, \boldsymbol{a}) - Q_i^{\pi_i^\dagger, \pi_{GA,-i}^*}(s, \boldsymbol{a}) \right|. \tag{4}$$

つまり，「抽象化したゲームと元のゲームにおけるミニマックス値の差」と，「元のゲームの均衡から抽象化したゲームの均衡へとプレイヤ $-i$ が逸脱したときにプレイヤ i が得られる期待利得のゲイン」によって抑えられる．それぞれの差の上界は，以下の補題によって与えられる．

補題 3.3. 任意の状態 $s \in \mathcal{S}$ と行動の組 $\boldsymbol{a} \in \mathcal{A}$ について次の不等式が成り立つ：

$$\left| Q_{A,i}^{\boldsymbol{\pi}_A^*}(\phi(s), \boldsymbol{a}) - Q_i^{\boldsymbol{\pi}^*}(s, \boldsymbol{a}) \right| \leq \frac{\epsilon}{1-\gamma}.$$

補題 3.4. 任意の状態 $s \in \mathcal{S}$ と行動の組 $\boldsymbol{a} \in \mathcal{A}$ について次の不等式が成り立つ：

$$\left| Q_i^{\boldsymbol{\pi}^*}(s, \boldsymbol{a}) - Q_i^{\pi_i^\dagger, \pi_{GA,-i}^*}(s, \boldsymbol{a}) \right| \leq \frac{2\epsilon}{(1-\gamma)^2}.$$

補題 3.3, 3.4と式 (4) より，$Q_{A,i}^{\boldsymbol{\pi}_A^*}(\phi(s), \boldsymbol{a})$ と $Q_i^{\pi_i^\dagger, \pi_{GA,-i}^*}(s, \boldsymbol{a})$ の差に関して以下の不等式を導ける：

$$\left| Q_{A,i}^{\boldsymbol{\pi}_A^*}(\phi(s), \boldsymbol{a}) - Q_i^{\pi_i^\dagger, \pi_{GA,-i}^*}(s, \boldsymbol{a}) \right| \leq \frac{3\epsilon}{(1-\gamma)^2}.$$

この不等式から，補題 3.2の仮定は $\delta = \frac{3\epsilon}{(1-\gamma)^2}$ で満たされることが分かる．したがって，補題 3.2から，

$$V_i^{\pi_i^\dagger, \pi_{GA,-i}^*}(s) - V_i^{\boldsymbol{\pi}_{GA}^*}(s) \leq \frac{6\epsilon}{(1-\gamma)^3}. \tag{5}$$

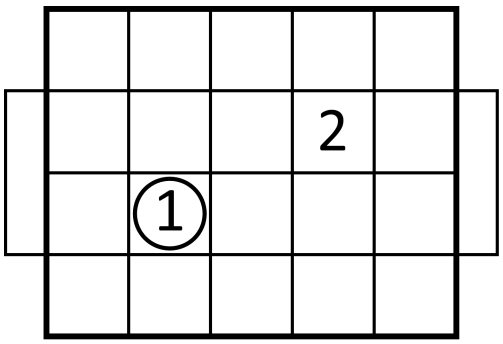


図 1: マルコフサッカーの初期状態（ボールの所持者がプレイヤ 1 の場合）

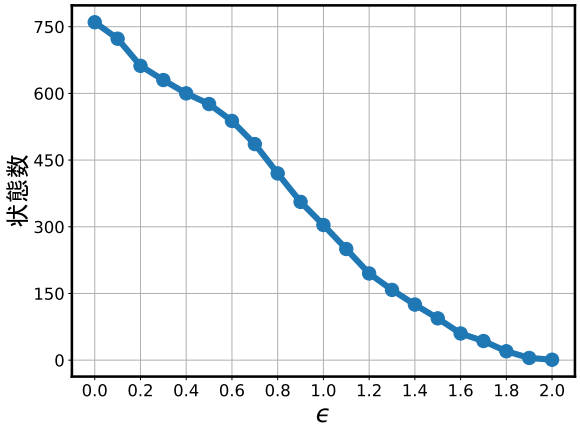


図 2: マルコフサッカーの状態数の削減の様子

最後に, 式 (3) と式 (5) より, 方策 $\boldsymbol{\pi}^*_{GA}$ についての Exploitability の上界は次のように導かれる：

$$\text{exploit}\,(\boldsymbol{\pi}^*_{GA}, s) \leq \sum_{i\in\{1,2\}} \frac{6\epsilon}{(1-\gamma)^3} = \frac{12\epsilon}{(1-\gamma)^3}.$$

□

4 計算機実験

本章では，二人零和マルコフゲームの一つであるマルコフサッカー [1, 7] に対して状態抽象化を適用し，計算機実験によってその性能を観察する．

4.1 実験設定

マルコフサッカーとは，二人のプレイヤによるシンプルなサッカーゲームである．本研究では，4×5 のフィールド上で行うマルコフサッカーについて考える．プレイヤの初期位置は図 1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．プレイヤがフィールド外のマスに移動する行動を選択した場合は，その行動のかわりに停止が実行される．選択された行動が実行される順番は確率によって決定される．プレイヤ同士が衝突する場合は，ボールの所持者はその場にとどまっていたプレイヤとなり，移動しようとしたプレイヤは移動前のマスにとどまる．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．図 1では，プレイヤ 1 は右側のゴールにゴールを決めると報酬 1 を得ることができ，反対にプレイヤ 2 は左側のゴールにゴールを決めると報酬 1 を得ることができる．また，初期状態からゴールが決まらずに 500 期経過した場合は引き分けとし，両者の報酬はともに 0 とする．

行動価値関数 $Q_i^{\boldsymbol{\pi}^*}$ に基づく状態抽象化をマルコフサッカーに適用し，状態数 $|\mathcal{S}_A|$ を計算した．ここで，状態抽象化の定義である式 (2) 中の ϵ を 0.0 から 2.0 まで 0.1 刻みで変化させ，各 ϵ の状態抽象化について確かめた．

元のマルコフサッカーの均衡方策 $\boldsymbol{\pi}^*$ や各 ϵ の状態抽象化を適用したマルコフサッカーの均衡方策 $\boldsymbol{\pi}^*_A$ をミニマックス Q 学習を用いて得た．ミニマックス Q 学習のパラメータは，学習の総ステップ数 L を $1,000,000$，割引因子 γ を 0.9 とした．また，学習率 α_t は学習ステップ数 $t \geq 0$ に対して $10^{-\frac{2}{L}t}$ とした．

プレイヤ 1 が方策 $\pi^*_{GA,1}$，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 ϵ ごとのプレイヤ 1 の勝率を求めた．ここで，各マルコフサッカーの実行回数は $10,000$ 回とし，各プレイヤの勝率は各プレイヤの報酬の総和を実行回数で割ったものとした．

最後に，元のマルコフサッカーの均衡方策 $\boldsymbol{\pi}^*$ や各 ϵ ごと状態抽象化を適用したマルコフサッカーの均衡方策 $\boldsymbol{\pi}^*_A$ について，学習ステップ数に対する Exploitability の推移を確認した．ここでは，ボールの所持者がプレイヤ 1 の場合の初期状態について計算した．また，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 [9].

4.2 実験結果

図 2 は，状態抽象化の適用による状態削減の様子を表している．横軸を ϵ，縦軸をマルコフサッカーの状態数としている．ϵ を 0.1 刻みで増加させると，状態抽象化が適用されたマルコフサッカーの状態数 $|\mathcal{S}_A|$ は線形に近い形で減少した．$\epsilon = 0.0$ での状態数は元のマルコフサッカーの状態数と同じ 760 であり，抽象化が行われなかったといえる．ゆえに，$\epsilon = 0.0$ におけるプレイヤ 1 の均衡方策 $\pi^*_{A,1}$ は元のマルコフサッカーの均衡方策 π^*_1 と同じであるといえる．また，マルコフサッカーの報酬は $-1, 0, 1$ であるため，同じ状態に抽象化される状態同士の行動価値関数の差は最大で 2 である．そのため，$\epsilon = 2.0$ における状態数は 1 であった．

図 3 は，プレイヤ 1 が方策 $\pi^*_{GA,1}$，プレイヤ 2 が方策 π^*_2 に従うマルコフサッカーにおけるプレイヤ 1 の勝率の推移を表している．横軸を ϵ，縦軸をプレイヤ 1 の勝率としている．二人のプレイヤがともに均衡方策 $\boldsymbol{\pi}^*$ に従うとき，つまり，$\epsilon = 0.0$ におけるプレイヤ 1 の勝率は 48.5% であった．同様

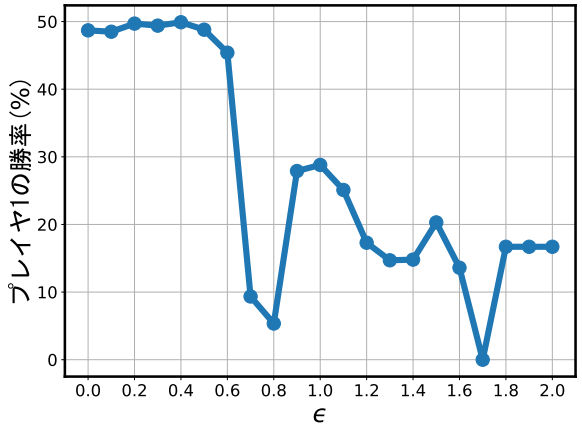


図 3: プレイヤ 1 の勝率の推移

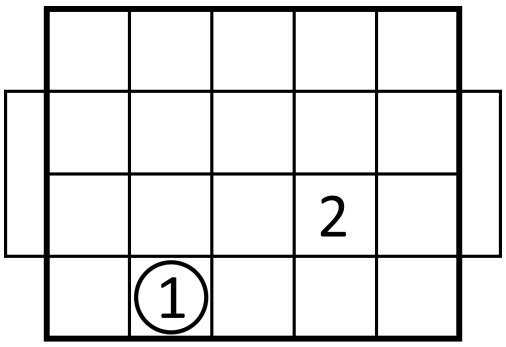


図 4: $\epsilon = 1.7$ のときに引き分けを起こす状態（状態 317）

に，$\epsilon = 0.6$ までの状態抽象化ならばプレイヤ 1 の勝率は 50% に近い値であったが，$\epsilon = 0.7, 0.8$ の状態抽象化では勝率が 10% 以下まで落ちた．そして，ϵ が 0.9 以上の状態抽象化を適用したときのプレイヤ 1 の勝率はほとんどが 10% から 30% の間におさまっており，$\epsilon = 1.7$ のときのみ 0% であった．

ϵ が $0.7, 0.8, 1.7$ の状態抽象化を適用したマルコフサッカーでは，引き分けを引き起こしやすい方策を学習していたため勝率が極端に低くなった．

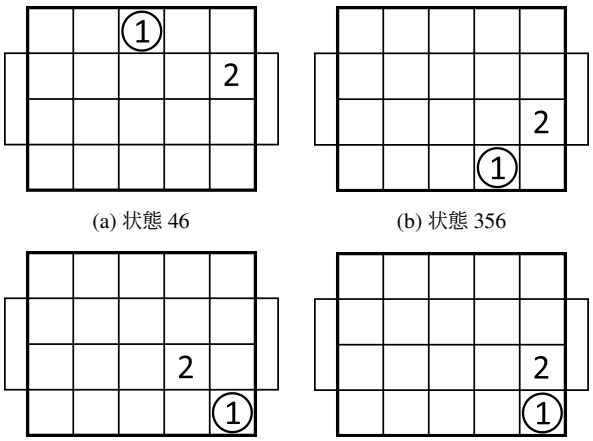
ϵ が 1.7 のときに勝率が 0% まで落ちる原因は，初期状態においてプレイヤ 1 がボールを所持している場合，二人のプレイヤが特定の位置から全く動かない状態（図 4）に陥りやすいからであった．ここではその状態を状態 317 と呼び，$\epsilon = 1.7$ の状態抽象化を状態 317 に適用したときの状態を抽象化状態 0 と呼ぶ．状態抽象化が適用されていないマルコフサッカーの均衡方策 $\boldsymbol{\pi}^*$ より，状態 317 ではプレイヤ 1 は 99.99% の確率で上への 1 マス移動を，プレイヤ 2 は 99.99% の確率でその場で停止の行動を選択する．しかし，$\epsilon = 1.7$ の状態抽象化が適用されたマルコフサッカーの均衡方策 $\boldsymbol{\pi}_A^*$ では，抽象化状態 0 においてプレイヤ 1 は 99.99% の確率で下への 1 マス移動を選択するように学習している．状態 317 においてプレイヤ 1 の 1 マス下には壁が存在するため，実際にはその場で停止する行動が実行される．ゆえに，$\epsilon = 1.7$ の状態抽象化を適用した場合は，状態 317 から別の状態に遷移するような行動が選択される確率が極めて低いといえる．また，$\epsilon = 1.7$ の状態抽象化が適用されたマルコフサッカーにおいて，プレイヤ 1 がボールを所持している場合の初期状態（図 1）から状態 317 に遷移する確率が極めて高い．ここではその初期状態を状態 217 と呼ぶ．$\epsilon = 1.7$ の状態抽象化を状態 217 に適用すると，状態 317 のときと同様に抽象化状態 0 に抽象化された．状態抽象化が適用されていないマルコフサッカーの均衡方策 $\boldsymbol{\pi}^*$ より，状態 217 ではプレイヤ 1 は 99.99% の確率で右への 1 マス移動を，プレイヤ 2 は 99.99% の確率で下への 1 マス移動の行動を選択する．しかし，抽象化状態 0 ではプレイヤ 1 は 99.99% の確率で下への 1 マス移動の行動が選択される．そのため，状態 217 では両プレイヤが下への 1 マス移動を実行し，状態 317 に遷移する確率が極めて高いといえる．ゆえに，引き分けの回数が増え，勝率が下がったといえる．

$\epsilon = 0.7, 0.8$ では，引き分けを引き起こすパターンが 2 種類存在した．ここでは，$\epsilon = 0.7$ について考える．1 つ目は，$\epsilon = 1.7$ と同様に，特定の状態から両プレイヤが全く動かないようなパターンであった（図 5a）．ここではその状態を状態 46 と呼ぶ．2 つ目は，特定の複数の状態間だけで遷移が繰り返されているパターンであった．ここではそれらの状態を状態 356, 374, 375 と呼ぶ（図 5b, 5c, 5d）．プレイヤ 1 が状態抽象化を適用したマルコフサッカーでの均衡方策 $\pi_{A,1}^*$，プレイヤ 2 が状態抽象化を適用していないマルコフサッカーでの均衡方策 π_2^* に従うマルコフサッカーについて考える．この場合，状態 356 は 99.95% の確率で状態 374 に遷移し，状態 374 は 86.93% の確率で状態 356 に遷移するように学習されていた．同様に，状態 374 は 13.06% の確率で状態 375 に遷移し，状態 375 は 99.95% の確率で状態 374 に遷移するように学習されていた．ゆえに，これら 3 つの状態から別の状態に遷移することは非常に困難であり，その結果引き分けが多発し勝率が下がったといえる．

図 6 は，学習ステップ数に対する Exploitability の推移を表している．ここで，横軸を学習ステップ数，縦軸を Exploitability とし，状態抽象化を適用していない元のマルコフサッカーの結果を”Ground” と表した．また，ϵ が $0.2, 0.6, 1.0, 1.4, 1.8$ についての状態抽象化を適用したマルコフサッカーの結果をプロットしている．プレイヤ 1 の勝率の結果と同様，ϵ が 0.6 までは Exploitability が 0 に近い値に収束した．

5 おわりに

本研究では，二人零和マルコフゲームの均衡計算の計算量を抑えるために，MDP の状態抽象化を二人零和マルコフゲームに拡張した．その性能の吟味のために，Exploitability の上界を導出した．今後は，行動価値関数に基づく状態抽象化以外の状態抽象化 [2] を二人零和マルコフゲームに拡張したり，



(a) 状態 46　(b) 状態 356

(c) 状態 374　(d) 状態 375

図 5: $\epsilon = 0.7$ のときに引き分けを引き起こす状態

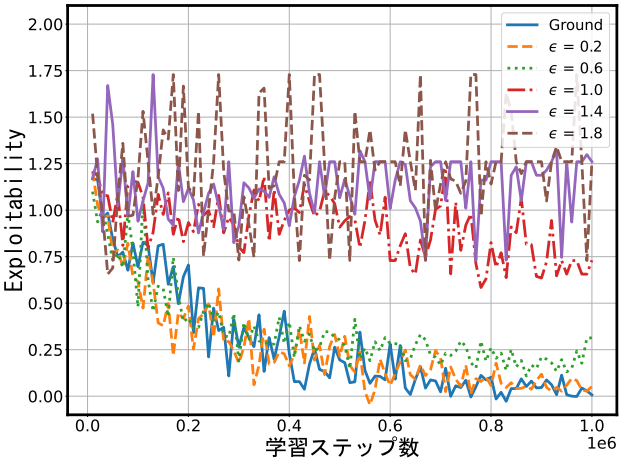


図 6: 学習ステップ数に対する Exploitability の推移

実際に規模の大きい二人零和マルコフゲームに今回の状態抽象化を導入し，その性能を評価したりしたい．

参考文献

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

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

[3] Pablo Samuel Castro. Scalable methods for computing state similarity in deterministic markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10069–10076, 2020.

[4] Norman Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite markov decision processes. *arXiv preprint arXiv:1207.4114*, 2012.

[5] Robert Givan, Thomas Dean, and Matthew Greig. Equivalence notions and model minimization in markov decision processes. *Artificial Intelligence*, 147(1-2):163–223, 2003.

[6] Lihong Li, Thomas J Walsh, and Michael L Littman. Towards a unified theory of state abstraction for mdps.

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

[8] Lloyd S Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10):1095–1100, 1953.

[9] 阿部拳之 and 金子雄祐. 二人零和マルコフゲームにおけるオフ方策評価のための q 学習.

A　定理 3.1 の補題の証明

A.1　補題 3.2 の証明

証明. まず，以下の補題を導入する：

補題 A.1. ある定数 $\delta \geq 0$ が存在し，任意の状態 $s \in \mathcal{S}$ と行動の組 $\boldsymbol{a} \in \mathcal{A}$ について，以下の式が成り立つとする：

$$\left| Q_i^{\pi_i^\dagger, \pi^*_{GA,-i}}(s, \boldsymbol{a}) - Q^{\boldsymbol{\pi}^*_A}_{A,i}(\phi(s), \boldsymbol{a}) \right| \leq \delta.$$

このとき，任意の状態 $s \in \mathcal{S}$ について次の不等式が成立する：

$$V_i^{\pi_i^\dagger, \pi^*_{GA,-i}}(s) \leq \sum_{\boldsymbol{a} \in \mathcal{A}} \boldsymbol{\pi}^*_{GA}(\boldsymbol{a}|s) Q_i^{\pi_i^\dagger, \pi^*_{GA,-i}}(s, \boldsymbol{a}) + 2\delta.$$

補題 A.1から，任意の $s \in \mathcal{S}$ に対して以下を得る：

$$\begin{aligned}
&V_i^{\pi_i^\dagger, \pi^*_{GA,-i}}(s) - V_i^{\boldsymbol{\pi}^*_{GA}}(s) \\
&\leq 2\delta + \sum_{\boldsymbol{a} \in \mathcal{A}} \boldsymbol{\pi}^*_{GA}(\boldsymbol{a}|s) \left(Q_i^{\pi_i^\dagger, \pi^*_{GA,-i}}(s, \boldsymbol{a}) - Q_i^{\boldsymbol{\pi}^*_{GA}}(s, \boldsymbol{a}) \right) \\
&= 2\delta + \gamma \sum_{\boldsymbol{a} \in \mathcal{A}} \boldsymbol{\pi}^*_{GA}(\boldsymbol{a}|s) \sum_{s' \in \mathcal{S}} P(s'|s, \boldsymbol{a}) \\
&\quad \cdot \left(V_i^{\pi_i^\dagger, \pi^*_{GA,-i}}(s') - V_i^{\boldsymbol{\pi}^*_{GA}}(s') \right) \\
&\leq 2\delta + \gamma \max_{s' \in \mathcal{S}} \left(V_i^{\pi_i^\dagger, \pi^*_{GA,-i}}(s') - V_i^{\boldsymbol{\pi}^*_{GA}}(s') \right).
\end{aligned}$$

両辺の最大値を取ると，

$$\begin{aligned}
&\max_{s' \in \mathcal{S}} \left(V_i^{\pi_i^\dagger, \pi^*_{GA,-i}}(s') - V_i^{\boldsymbol{\pi}^*_{GA}}(s') \right) \\
&\leq 2\delta + \gamma \max_{s' \in \mathcal{S}} \left(V_i^{\pi_i^\dagger, \pi^*_{GA,-i}}(s') - V_i^{\boldsymbol{\pi}^*_{GA}}(s') \right).
\end{aligned}$$

最後に，右辺の $\gamma \max_{s' \in \mathcal{S}} \left(V_i^{\pi_i^\dagger, \pi^*_{GA,-i}}(s') - V_i^{\boldsymbol{\pi}^*_{GA}}(s') \right)$ を左辺に移項することで，補題の主張を示すことができる：

$$\begin{aligned}
&V_i^{\pi_i^\dagger, \pi^*_{GA,-i}}(s) - V_i^{\boldsymbol{\pi}^*_{GA}}(s) \\
&\leq \max_{s' \in \mathcal{S}} \left(V_i^{\pi_i^\dagger, \pi^*_{GA,-i}}(s') - V_i^{\boldsymbol{\pi}^*_{GA}}(s') \right) \leq \frac{2\delta}{1-\gamma}.
\end{aligned}$$

□

A.2 補題 A.1 の証明

証明. $V_i^{\pi_i^\dagger,\pi^*_{GA,-i}}$ の定義は以下のとおりである：

$$
\begin{aligned}
&V_i^{\pi_i^\dagger,\pi^*_{GA,-i}}(s)\\
&=\sum_{\boldsymbol{a}\in\mathcal{A}}\pi_i^\dagger(a_i|s)\pi^*_{GA,-i}(a_{-i}|s)Q_i^{\pi_i^\dagger,\pi^*_{GA,-i}}(s,\boldsymbol{a}).
\end{aligned}
$$

仮定より，

$$
\begin{aligned}
&V_i^{\pi_i^\dagger,\pi^*_{GA,-i}}(s)\\
&\le\sum_{\boldsymbol{a}\in\mathcal{A}}\pi_i^\dagger(a_i|s)\pi^*_{GA,-i}(a_{-i}|s)Q_{A,i}^{\pi^*_A}(\phi(s),\boldsymbol{a})+\delta\\
&=\sum_{\boldsymbol{a}\in\mathcal{A}}\pi_i^\dagger(a_i|s)\pi^*_{A,-i}(a_{-i}|\phi(s))Q_{A,i}^{\pi^*_A}(\phi(s),\boldsymbol{a})+\delta\\
&\le\sum_{\boldsymbol{a}\in\mathcal{A}}\pi^*_{A,i}(a_i|\phi(s))\pi^*_{A,-i}(a_{-i}|\phi(s))Q_{A,i}^{\pi^*_A}(\phi(s),\boldsymbol{a})+\delta\\
&\le\sum_{\boldsymbol{a}\in\mathcal{A}}\pi^*_{A,i}(a_i|\phi(s))\pi^*_{A,-i}(a_{-i}|\phi(s))Q_i^{\pi_i^\dagger,\pi^*_{GA,-i}}(s,\boldsymbol{a})+2\delta\\
&=\sum_{\boldsymbol{a}\in\mathcal{A}}\pi^*_{GA,i}(a_i|s)\pi^*_{GA,-i}(a_{-i}|s)Q_i^{\pi_i^\dagger,\pi^*_{GA,-i}}(s,\boldsymbol{a})+2\delta.
\end{aligned}
$$

□

A.3 補題 3.3 の証明

証明. 式 (1) におけるミニマックス値の定義より，任意の $s\in\mathcal{S}$ および $\boldsymbol{a}\in\mathcal{A}$ に対して，

$$
\begin{aligned}
&\sum_{g\in G(s)}w(g)Q_i^{\boldsymbol{\pi}^*}(g,\boldsymbol{a})-Q_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s),\boldsymbol{a})\\
&=\gamma\sum_{g\in G(s)}w(g)\sum_{s'\in\mathcal{S}}P(s'|g,\boldsymbol{a})\\
&\quad\cdot\left(\max_{p\in\Delta(A_i)}\min_{a'_{-i}\in\mathcal{A}_{-i}}\sum_{a'_i\in\mathcal{A}_i}p(a'_i)Q_i^{\boldsymbol{\pi}^*}(s',\boldsymbol{a}')\right.\\
&\quad\left.-\max_{p\in\Delta(A_i)}\min_{a'_{-i}\in\mathcal{A}_{-i}}\sum_{a'_i\in\mathcal{A}_i}p(a'_i)Q_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s'),\boldsymbol{a}')\right)\\
&\le\gamma\max_{(s',\boldsymbol{a}')\in\mathcal{S}\times\mathcal{A}}\left(Q_i^{\boldsymbol{\pi}^*}(s',\boldsymbol{a}')-Q_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s'),\boldsymbol{a}')\right).
\end{aligned}
$$

したがって，仮定 (2) より，

$$
\begin{aligned}
&\gamma\max_{(s',\boldsymbol{a}')\in\mathcal{S}\times\mathcal{A}}\left(Q_i^{\boldsymbol{\pi}^*}(s',\boldsymbol{a}')-Q_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s'),\boldsymbol{a}')\right)\\
&\ge\sum_{g\in G(s)}w(g)Q_i^{\boldsymbol{\pi}^*}(g,\boldsymbol{a})-Q_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s),\boldsymbol{a})\\
&\ge\min_{g\in G(s)}Q_i^{\boldsymbol{\pi}^*}(g,\boldsymbol{a})-Q_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s),\boldsymbol{a})\\
&\ge-\epsilon+Q_i^{\boldsymbol{\pi}^*}(s,\boldsymbol{a})-Q_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s),\boldsymbol{a}).
\end{aligned}
$$

両辺の最大値を取ると，

$$
\begin{aligned}
&\max_{(s,\boldsymbol{a})\in\mathcal{S}\times\mathcal{A}}\left(Q_i^{\boldsymbol{\pi}^*}(s,\boldsymbol{a})-Q_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s),\boldsymbol{a})\right)\\
&\le\epsilon+\gamma\max_{(s,\boldsymbol{a})\in\mathcal{S}\times\mathcal{A}}\left(Q_i^{\boldsymbol{\pi}^*}(s,\boldsymbol{a})-Q_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s),\boldsymbol{a})\right).
\end{aligned}
$$

この不等式を整理することで，以下を得る：

$$
\begin{aligned}
&Q_i^{\pi^*}(s,\boldsymbol{a})-Q_{A,i}^{\pi^*_A}(\phi(s),\boldsymbol{a})\\
&\le\max_{(s',\boldsymbol{a}')\in\mathcal{S}\times\mathcal{A}}\left(Q_i^{\pi^*}(s',\boldsymbol{a}')-Q_{A,i}^{\pi^*_A}(\phi(s'),\boldsymbol{a}')\right)\\
&\le\frac{\epsilon}{1-\gamma}.
\end{aligned}
$$

同様に，任意の $s\in\mathcal{S}$ および $\boldsymbol{a}\in\mathcal{A}$ に対して，

$$
\begin{aligned}
&\sum_{g\in G(s)}w(g)Q_i^{\boldsymbol{\pi}^*}(g,\boldsymbol{a})-Q_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s),\boldsymbol{a})\\
&\ge\gamma\min_{(s',\boldsymbol{a}')\in\mathcal{S}\times\mathcal{A}}\left(Q_i^{\boldsymbol{\pi}^*}(s',\boldsymbol{a}')-Q_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s'),\boldsymbol{a}')\right).
\end{aligned}
$$

よって，仮定 (2) より，

$$
\begin{aligned}
&\gamma\min_{(s',\boldsymbol{a}')\in\mathcal{S}\times\mathcal{A}}\left(Q_i^{\boldsymbol{\pi}^*}(s',\boldsymbol{a}')-Q_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s'),\boldsymbol{a}')\right)\\
&\le\sum_{g\in G(s)}w(g)Q_i^{\boldsymbol{\pi}^*}(g,\boldsymbol{a})-Q_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s),\boldsymbol{a})\\
&\le\max_{g\in G(s)}Q_i^{\boldsymbol{\pi}^*}(g,\boldsymbol{a})-Q_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s),\boldsymbol{a})\\
&\le\epsilon+Q_i^{\boldsymbol{\pi}^*}(s,\boldsymbol{a})-Q_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s),\boldsymbol{a}).
\end{aligned}
$$

両辺の最小値を取ると，

$$
\begin{aligned}
&\min_{(s,\boldsymbol{a})\in\mathcal{S}\times\mathcal{A}}\left(Q_i^{\boldsymbol{\pi}^*}(s,\boldsymbol{a})-Q_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s),\boldsymbol{a})\right)\\
&\ge-\epsilon+\gamma\min_{(s,\boldsymbol{a})\in\mathcal{S}\times\mathcal{A}}\left(Q_i^{\boldsymbol{\pi}^*}(s,\boldsymbol{a})-Q_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s),\boldsymbol{a})\right).
\end{aligned}
$$

よって，

$$
\begin{aligned}
&Q_i^{\boldsymbol{\pi}^*}(s,\boldsymbol{a})-Q_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s),\boldsymbol{a})\\
&\ge\min_{(s',\boldsymbol{a}')\in\mathcal{S}\times\mathcal{A}}\left(Q_i^{\boldsymbol{\pi}^*}(s',\boldsymbol{a}')-Q_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s'),\boldsymbol{a}')\right)\\
&\ge-\frac{\epsilon}{1-\gamma}.
\end{aligned}
$$

以上をまとめると，補題の主張を示すことができる：

$$
\left|Q_i^{\boldsymbol{\pi}^*}(s,\boldsymbol{a})-Q_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s),\boldsymbol{a})\right|\le\frac{\epsilon}{1-\gamma}.
$$

□

A.4 補題 3.4 の証明

証明. 行動価値関数の定義から，任意の $s\in\mathcal{S}$ および $\boldsymbol{a}\in\mathcal{A}$ に対して，

$$
\begin{aligned}
&Q_i^{\pi_i^\dagger,\pi^*_{GA,-i}}(s,\boldsymbol{a})-Q_i^{\boldsymbol{\pi}^*}(s,\boldsymbol{a})\\
&=\gamma\sum_{s'\in\mathcal{S}}P(s'|s,\boldsymbol{a})\left(V_i^{\pi_i^\dagger,\pi^*_{GA,-i}}(s')-V_i^{\boldsymbol{\pi}^*}(s')\right).
\end{aligned}
$$

ここで，$V_i^{\boldsymbol{\pi}^*}$ の上・下界に関する以下の補題を導入する：

補題 A.2. 任意の状態 $s\in\mathcal{S}$ について，以下の不等式が成り立つ：

$$
\left|V_i^{\boldsymbol{\pi}^*}(s)-V_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s))\right|\le\frac{\epsilon}{1-\gamma}.
$$

補題 A.2を適用すると，以下の不等式を得る：

$$
\begin{aligned}
&Q_i^{\pi_i^\dagger,\pi^*_{GA,-i}}(s,\boldsymbol{a}) - Q_i^{\boldsymbol{\pi}^*}(s,\boldsymbol{a}) \\
&\leq \frac{\gamma\epsilon}{1-\gamma} \\
&\quad + \gamma \sum_{s'\in\mathcal{S}} P(s'|s,\boldsymbol{a}) \left(V_i^{\pi_i^\dagger,\pi^*_{GA,-i}}(s') - V_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s')) \right) \\
&= \frac{\gamma\epsilon}{1-\gamma} + \gamma \sum_{s'\in\mathcal{S}} P(s'|s,\boldsymbol{a}) \\
&\quad \cdot \left(\max_{a'_i\in\mathcal{A}_i} \sum_{a'_{-i}\in\mathcal{A}_{-i}} \pi^*_{GA,-i}(a'_{-i}|s') Q_i^{\pi_i^\dagger,\pi^*_{GA,-i}}(s',\boldsymbol{a}') \right. \\
&\quad \left. - \max_{a'_i\in\mathcal{A}_i} \sum_{a'_{-i}\in\mathcal{A}_{-i}} \pi^*_{GA,-i}(a'_{-i}|s') Q_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s'),\boldsymbol{a}') \right) \\
&\leq \frac{\gamma\epsilon}{1-\gamma} + \gamma \sum_{s'\in\mathcal{S}} P(s'|s,\boldsymbol{a}) \max_{a'_i\in\mathcal{A}_i} \sum_{a'_{-i}\in\mathcal{A}_{-i}} \pi^*_{GA,-i}(a'_{-i}|s') \\
&\quad \cdot \left(Q_i^{\pi_i^\dagger,\pi^*_{GA,-i}}(s',\boldsymbol{a}') - Q_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s'),\boldsymbol{a}') \right) \\
&\leq \frac{\gamma\epsilon}{1-\gamma} \\
&\quad + \gamma \max_{(s',\boldsymbol{a}')\in\mathcal{S}\times\mathcal{A}} \left(Q_i^{\pi_i^\dagger,\pi^*_{GA,-i}}(s',\boldsymbol{a}') - Q_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s'),\boldsymbol{a}') \right).
\end{aligned}
$$

よって，補題 3.3から，任意の $s\in\mathcal{S}$ および $\boldsymbol{a}\in\mathcal{A}_1\times\mathcal{A}_2$ に対して，

$$
\begin{aligned}
&Q_i^{\pi_i^\dagger,\pi^*_{GA,-i}}(s,\boldsymbol{a}) - Q_i^{\boldsymbol{\pi}^*}(s,\boldsymbol{a}) \leq \frac{2\gamma\epsilon}{1-\gamma} \\
&\quad + \gamma \max_{(s',\boldsymbol{a}')\in\mathcal{S}\times\mathcal{A}} \left(Q_i^{\pi_i^\dagger,\pi^*_{GA,-i}}(s',\boldsymbol{a}') - Q_i^{\boldsymbol{\pi}^*}(s',\boldsymbol{a}') \right).
\end{aligned}
$$

両辺の最大値を取ると，

$$
\begin{aligned}
&\max_{(s,\boldsymbol{a})\in\mathcal{S}\times\mathcal{A}} \left(Q_i^{\pi_i^\dagger,\pi^*_{GA,-i}}(s,a_1,a_2) - Q_i^{\boldsymbol{\pi}^*}(s,\boldsymbol{a}) \right) \\
&\leq \frac{2\gamma\epsilon}{1-\gamma} + \gamma \max_{(s,\boldsymbol{a})\in\mathcal{S}\times\mathcal{A}} \left(Q_i^{\pi_i^\dagger,\pi^*_{GA,-i}}(s,\boldsymbol{a}) - Q_i^{\boldsymbol{\pi}^*}(s,\boldsymbol{a}) \right).
\end{aligned}
$$

よって，右辺の $\gamma\max_{(s,\boldsymbol{a})\in\mathcal{S}\times\mathcal{A}} \left(Q_i^{\pi_i^\dagger,\pi^*_{GA,-i}}(s,\boldsymbol{a}) - Q_i^{\boldsymbol{\pi}^*}(s,\boldsymbol{a}) \right)$ を左辺に移項することで，以下の不等式を得る：

$$
\begin{aligned}
&Q_i^{\pi_i^\dagger,\pi^*_{GA,-i}}(s,\boldsymbol{a}) - Q_i^{\boldsymbol{\pi}^*}(s,\boldsymbol{a}) \\
&\leq \max_{(s',\boldsymbol{a}')\in\mathcal{S}\times\mathcal{A}} \left(Q_i^{\pi_i^\dagger,\pi^*_{GA,-i}}(s',\boldsymbol{a}') - Q_i^{\boldsymbol{\pi}^*}(s',\boldsymbol{a}') \right) \\
&\leq \frac{2\epsilon}{(1-\gamma)^2}. \qquad (6)
\end{aligned}
$$

一方，ミニマックス値の性質から，

$$
Q_i^{\boldsymbol{\pi}^*}(s,\boldsymbol{a}) - Q_i^{\pi_i^\dagger,\pi^*_{GA,-i}}(s,\boldsymbol{a}) \leq 0. \qquad (7)
$$

式 (6)，(7) を組み合わせると，補題の主張を示すことができる：

$$
\left| Q_i^{\pi_i^\dagger,\pi^*_{GA,-i}}(s,\boldsymbol{a}) - Q_i^{\boldsymbol{\pi}^*}(s,\boldsymbol{a}) \right| \leq \frac{2\epsilon}{(1-\gamma)^2}.
$$

□

A.5 補題 A.2 の証明

証明. $V_{A,i}^{\boldsymbol{\pi}^*_A}$ の定義より，任意の状態 $s\in\mathcal{S}$ に対して，

$$
V_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s)) = \max_{p\in\Delta(A_i)} \min_{a_{-i}\in\mathcal{A}_{-i}} \sum_{a_i\in\mathcal{A}_i} p(a_i) Q_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s),\boldsymbol{a}).
$$

この式の $Q_{A,i}^{\boldsymbol{\pi}^*_A}$ に補題 3.3 を適用すると，以下の不等式を得る：

$$
\begin{aligned}
&V_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s)) \\
&\leq \max_{p\in\Delta(A_i)} \min_{a_{-i}\in\mathcal{A}_{-i}} \sum_{a_i\in\mathcal{A}_i} p(a_i) Q_i^{\boldsymbol{\pi}^*}(s,\boldsymbol{a}) + \frac{\epsilon}{1-\gamma} \\
&= V_i^{\boldsymbol{\pi}^*}(s) + \frac{\epsilon}{1-\gamma}.
\end{aligned}
$$

同様に，

$$
\begin{aligned}
&V_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s)) \\
&\geq \max_{p\in\Delta(A_i)} \min_{a_{-i}\in\mathcal{A}_{-i}} \sum_{a_i\in\mathcal{A}_i} p(a_i) Q_i^{\boldsymbol{\pi}^*}(s,\boldsymbol{a}) - \frac{\epsilon}{1-\gamma} \\
&= V_i^{\boldsymbol{\pi}^*}(s) - \frac{\epsilon}{1-\gamma}.
\end{aligned}
$$

したがって，二つの不等式を組み合わせると，

$$
\left| V_i^{\boldsymbol{\pi}^*}(s) - V_{A,i}^{\boldsymbol{\pi}^*_A}(\phi(s)) \right| \leq \frac{\epsilon}{1-\gamma}.
$$

□