「機械学習の制御技術への活用」論文特集号—II

論　文

マルチエージェントシステムのための深層強化学習における効率的データ共有*

林田　智弘†・浅野光太郎†・関崎　真也†・西崎　一郎†

Data Sharing on Deep Reinforcement Learning for Multi-Agent Systems*

Tomohiro Hayashida†, Kotaro Asano†, Shinya Sekizaki† and Ichiro Nishizaki†

In recent years, reinforcement learning has advanced in various fields such as autonomous driving, behavior analysis in power markets, and robot control. These studies have demonstrated the utility of applying MAS (Multi-Agent System), where multiple agents cooperate and compete with each other. In the learning process within MAS, a significant challenge is that irregular fluctuations in the environment caused by the actions of other agents increase environmental uncertainty and destabilize learning. To overcome this challenge, research has been reported on methods to improve learning efficiency by effectively utilizing the experience data of other agents. Previous studies have reported cases where sharing all data improved the learning efficiency of agents in cooperative or competitive relationships, as well as cases where sharing only a portion of the experience data improved learning efficiency. This paper aims to improve learning efficiency in MAS by emphasizing the similarity between agents as a criterion for data sharing and demonstrating its effectiveness. Specifically, we propose a new method that selects shared data based on the similarity between agents and treats it as experience data that complements one's own experience. Furthermore, we demonstrate that the proposed method improves the learning efficiency of agents through simulations involving asymmetric agents.

1. はじめに

近年，強化学習の発展により，自動運転技術，電力市場における需給調整やロボット協調制御など多岐にわたる応用が進展している．これらの分野では，複数のエージェントが相互作用しながら行動する環境が重要であり，マルチエージェントシステム (MAS: Multi-Agent System) の採用が注目されている．特にMAS環境では，エージェント間の相互作用が環境の不確実性を増大させ，学習プロセスが不安定化するという課題が存在する．本論文ではこの課題に対応するために，エージェント間の類似性を基にしたデータ共有戦略を提案する．提案手法はMAS環境における学習効率を向上させ，異なる特性を持つエージェント間での適応的な行動選択を実現するものである．これによりMASの学習プロセスが安定化し，強化学習を活用した幅広い制御技術や情報システムの基盤技術としての応用が期待される．

Shwartz *et al.*[1] は，MASを想定した自動運転において，安全な運転行動を獲得するための強化学習手法を提案した．Ye *et al.*[2] は，MASを活用した電力市場における経済主体の戦略的入札行動の学習手法を提案した．Namalomba *et al.*[3] は，電気料金に対し弾力的な行動選択を行う需要家が存在する電力市場において，発電事業者の戦略的な入札をMASを用いて分析した．Orr *et al.*[4] は，複数のロボットを使用したマルチエージェント深層強化学習 (MADRL: Multi-Agent Deep Reinforcement Learning) について調査し，複数のロボットが共同でタスクを達成するシナリオに焦点を当て，様々な応用分野における最新の研究動向と課題を紹介している．

シングルエージェントシステム (SAS: Single-Agent Systems) と比較して，MASはエージェント同士の相互作用により環境変動が増し，不確実性が増大しやすく，学習プロセスが不安定化しやすい [5]．このため，協調や競争行動が発生するマルチエージェント環境では，従

* 原稿受付　2024年8月2日
* 広島大学 大学院 先進理工系科学研究科　Graduate School of Advanced Science and Engineering, Hiroshima University; 1-4-1, Kagamiyama, Higashihiroshima, Hiroshima 739-8527, JAPAN
Key Words: reinforcement learning, multi-agent systems, data sharing, learning efficiency.

来の学習アルゴリズム（例：ディープ Q ネットワーク (DQN: Deep Q-Network)[6] やディープ確定的方策勾配法 (DDPG: Deep Deterministic Policy Gradient)[7]）では安定した学習が難しい．この課題を克服するため，他のエージェントの経験データを活用し，学習効率を向上させ，不確実性を抑制する MADRL が多数提案されている [5,8–10]．Lowe *et al.*[5] は，アクター・クリティック手法に基づき，全エージェントの観測と行動の履歴を共有して協調および競争的な行動を学習する手法を提案した．Rashid *et al.*[8] は，DQN と長短期記憶 (LSTM: Long Short-Term Memory) を組み合わせた深層再帰型 Q 学習ネットワーク (DRQN: Deep Recurrent Q-Learning Network) を MAS に適用し，Mixing Network を通じて環境全体の Q 値を推定する手法を提案した．Hayashida *et al.*[9] は，すべてのデータを共有するのではなく，一部の経験データのみを共有することで学習効率が改善される手法を提案した．Gerstgrasser *et al.*[10] は，優先経験再生 (PER: Prioritized Experience Replay) に基づき，TD 誤差の大きいデータを優先的に共有する手法を提案し，DQN と比較して高い報酬を獲得した．

本論文の目的は，エージェント間での効果的な経験データ共有を通じて，MAS における学習効率と精度を向上させることである．とくに，異なる能力や報酬を持つエージェントが混在する非対称な環境に焦点を当て，類似性の高いエージェント同士でデータを共有し，異なる特徴を持つエージェントとはあまり共有しない手法を提案する．また，学習進捗に基づいてデータ共有率を調整し，学習初期段階での不適切なデータ共有を抑制する．さらに，累積報酬や行動傾向に基づく 2 種類のデータ選択方法を提案し，それぞれの方法で学習効率を向上させる．

本論文の構成は次の通りである．**2** 節で強化学習や MAS に関する知識と先行研究を紹介し，**3** 節でエージェント間の類似性指標の構築とそれに基づく経験データ共有の学習手法を提案する．**4** 節で，非対称なエージェントを用いたシミュレーション実験を実施し，提案手法の学習効率向上を示す．最後に，**5** 節で本論文の総括と今後の課題を述べる．

2. 深層強化学習とマルチエージェントシステム

2.1 深層強化学習

深層強化学習 (DRL: Deep Reinforcement Learning) は，深層ニューラルネットワーク (DNN: Deep Neural Network) を強化学習に応用した手法であり，画像識別や機械翻訳などで使用される [11]．DNN には，順伝播型ニューラルネットワーク，畳み込みニューラルネットワーク，再帰型ニューラルネットワークなどがある [12]．代表的な DRL の一種に，Q 学習に DNN を導入した Deep Q-Network (DQN) があり，Q 学習における次元の呪いを解消して複雑な問題の解決を可能にした [6,9]．

2.2 アクター・クリティック

アクター・クリティック手法は，強化学習における代表的な手法の一種であり，方策勾配法と価値関数法の利点を組み合わせたものである．本手法では，行動を選択するアクターと選択された行動の評価を行うクリティックの二つの構成要素が相互に連携して学習を進める．アクター・クリティックの基本概念は人間の意思決定とその評価にたとえることができる．たとえば，アクターにより行動選択の役割を担い，クリティックは選択した行動を評価する役割を持つことで，次の選択がより良いものになるようにアクターおよびクリティックが調整される．具体的には，アクターは方策 $\pi(a_t \mid \mathbf{o}_t;\phi)$ を定義し，観測 $\mathbf{o}_t$ に基づいて行動 a_t を選択する．一方で，クリティックは状態価値関数 $V(\mathbf{o}_t \mid \varphi)$ を用いて，選択した行動の環境に対する影響を定量的に評価する．このように，アクターはクリティックからの評価を基に方策を更新することで，行動の選択を改善する．すなわち，アクター・クリティックは行動選択とその評価を独立して学習を行うため，学習の効率化と安定性の向上が期待できる．とくに，連続的な行動空間を扱う場面や，複雑な環境での適応的な方策学習において有効である．学習時には TD 誤差を算出し，アクターとクリティックのパラメータ $\theta=(\phi,\varphi)$ を最適化する．(1), (2) 式は，それぞれアクターとクリティックの損失関数を示し，最小化することで行動の改善と行動価値関数の精度向上を目的とする．

$$\mathcal{L}(\phi) = -\log\pi(a_t \mid o_t;\phi) \\ (r_t+\gamma V(o_{t+1};\varphi)-V(o_t;\varphi)) \tag{1}$$

$$\mathcal{L}(\varphi) = \{r_t+\gamma V(o_{t+1};\varphi)-V(o_t;\varphi)\}^2 \tag{2}$$

2.3 エージェント間の情報共有を考慮したマルチエージェントシステム

Matignon *et al.*[14] は独立して Q 学習 [13] を用いて方策獲得するエージェントが，いくつかの MAS において学習に失敗したことを報告した．MAS の環境の不安定性により経験再生が正常に機能しなかったことが原因と考えられる [5]．

2.3.1 MADDPG

Multi-Agent Deep Deterministic Policy Gradient (MADDPG)[5] はアクター・クリティック構造を持つ DQN の一種で，DDPG (Deep Deterministic Policy Gradient)[7] の拡張モデルである．MADDPG では，エージェントがすべてのデータを共有することで協調や競争行動を学習する．各エージェントは，自身の観測を入力として行動を決定するアクターネットワークと，全エージェントの観測と行動を入力して価値推定を行うクリティックネットワークを持つ．

2.3.2 QMIX

QMIX[8] は，自律的に行動および学習する複数のエージェントが，一括して集中的に学習を行うCTDE (Centralized Training with Decentralized Execution) の一種であり，DQNとLSTMを組み合わせたDRQNを拡張した手法である．各エージェントは観測履歴と直近の行動を用いてQ値を計算し，Mixing Networkを通じて環境全体のQ値を学習する．Mixing Networkは，入力を単調に混合して価値関数を近似し，エージェントの個別Q値が全体の最適行動選択に貢献するように設計されている．

2.3.3 経験データの部分的共有による学習効率の向上

Hayashida *et al.*[9] は，すべての経験データを共有することが学習効率に必ずしも有用ではないとし，部分的にデータを共有する三つの手法を提案している．

ランダム方式: 経験データをランダムに抽出して共有する方法

固定割合方式: 経験データを累積報酬に基づいて並べ替え，上位と下位のデータを固定割合で共有する方法

変動割合方式: 学習の進行に応じて共有するデータの割合を適応的に調整する方法

これらの手法では，各エージェントが独立したアクター・クリティックを用いるためのネットワークを持つ．エピソードごとに経験データを記録し，一部を共有することで，計算時間を抑制しつつ学習効率の向上が期待される．上記の三つの手法によって選択された他エージェントから共有されたデータは，自身の経験データと同様に学習に用いられる．シミュレーション実験により，すべてのデータを共有するのではなく，部分的にデータを共有することで学習効率が向上することが示されている．

3. 効率的な学習データの共有方法の提案

本論文では，MASにおけるエージェントの学習効率向上を実現するため，エージェント間の類似性に基づく効率的なデータ共有手法を提案する．ここで，エージェントの類似性は，同一または類似した観測におけるエージェントの行動分布に基づいて評価される．類似性が高いエージェント同士は多くのデータを共有し，類似性が低い場合はほとんどデータを共有しない．このようなデータ共有の仕組みにより，各エージェントの特性を考慮した効果的なデータ共有が可能となり，学習効率の向上が期待される．

3.1 エージェント間の類似度

エージェント i の観測空間を O_i，各エピソードの最大ステップ数を T とし，エージェント i のエピソード k の観測履歴を $H_{i,k} \equiv \{o_{i,k,t} \mid o_{i,k,t} \in O_i, 1 \le t \le T\}$ と定義する．任意のエピソード e におけるエージェント i の直近 EP_A エピソード分の観測履歴の集合を $H_i^{EP_A} \equiv \{H_{i,ep} \mid \max\{e-EP_A+1,0\} \le ep \le e\}$ と定義する．エージェントの組 (i,j) を考えたとき，エージェント i,j の観測履歴 $H_i^{EP_A}$，$H_j^{EP_A}$ からそれぞれランダムに EP_B $(EP_A \ge EP_B)$ エピソード分の観測履歴 $H_i^{EP_A,EP_B} \subseteq H_i^{EP_A}$，$H_j^{EP_A,EP_B} \subseteq H_j^{EP_A}$ を選択し，要素の重複のない観測集合 $O_i^{EP_A,EP_B} \subseteq H_i^{EP_A,EP_B}$，$O_j^{EP_A,EP_B} \subseteq H_j^{EP_A,EP_B}$ を生成する．

任意の観測ベクトル $\mathbf{o}_i = \{o_{i,1}, o_{i,2}, \ldots\} \in O_i^{EP_A,EP_B}$，$\mathbf{o}_j = \{o_{j,1}, o_{j,2}, \ldots\} \in O_j^{EP_A,EP_B}$ の各要素 $o_{i,m}$，$o_{j,n}$ を0-1正規化したベクトル o'_i，o'_j の $L2$ ノルム距離を観測ベクトル $\mathbf{o}_i$，$\mathbf{o}_j$ の距離 $d(\mathbf{o}_i, \mathbf{o}_j)$ とし，観測ベクトル間の距離について昇順に最大 SMo 個の観測組の集合をエージェント i，j の類似する観測組集合 $\mathcal{P}_{i,j}$ と定義する．つぎに，$\mathcal{P}_{i,j}$ に内包される観測の組 $\mathbf{o}_i, \mathbf{o}_j \in \mathcal{P}_{i,j}$ を用いて，方策分布の非類似度を計算する．観測 $\mathbf{o}_i$,$\mathbf{o}_j$ に対するエージェント i,j の方策分布 $\pi_{\theta_i}(\mathbf{o}_i)$，$\pi_{\theta_j}(\mathbf{o}_j)$ について，(3) 式に従い全変動距離 $tv(\pi_{\theta_i}(\mathbf{o}_i), \pi_{\theta_j}(\mathbf{o}_j))$ を測定し，観測の組 $\mathbf{o}_i$,$\mathbf{o}_j$ におけるそれぞれのエージェントの方策の距離とする．

$$tv(\pi^{\theta_i}(\mathbf{o}_i), \pi^{\theta_j}(\mathbf{o}_j)) = \frac{1}{2} \sum_{a \in \mathcal{A}} |\pi^{\theta_i}(a|\mathbf{o}_i) - \pi^{\theta_j}(a|\mathbf{o}_j)| \tag{3}$$

本論文では，すべてのエージェントが共通する行動空間を $\mathcal{A}$ とし，集合 $\mathcal{P}_{i,j}$ の要素についてエージェントの方策間の距離に基づく非類似度 $u_{i,j}^{\mathrm{DS}}$ を次式で定義する．

$$u_{i,j}^{\mathrm{DS}} = \frac{1}{|\mathcal{P}_{i,j}|} \sum_{(o_i, o_j) \in \mathcal{P}_{i,j}} (d(o_i, o_j) + \epsilon) \times ((tv(\pi^{\theta_i}(o_i), \pi^{\theta_j}(o_j)) + \epsilon)) \tag{4}$$

ここで，$\epsilon > 0$ は十分に小さい正の実数である．

3.2 未学習度

エージェントの学習が不十分な段階では，非類似度による特徴分類が難しいため，本論文では未学習度を導入してデータ共有割合を調整する．エージェントが学習を進めると特定の行動を選択し，方策分布 π_θ のエントロピー $H(\pi_\theta(\mathbf{o}))$ が0に近づく．エージェント i と j のエントロピーの積として未学習度 $u_{i,j}^{UL}$ を次式で定義する．

$$u_{i,j}^{UL} = \frac{1}{|\mathcal{P}_{i,j}|} \sum_{(\mathbf{o}_i, \mathbf{o}_j) \in \mathcal{P}_{i,j}} (H(\pi^{\theta_i}(\mathbf{o}_i)) \times H(\pi^{\theta_j}(\mathbf{o}_j)))$$
$$H(\pi^\theta(\mathbf{o})) = \sum_{a \in \mathcal{A}} \pi^\theta(a, \mathbf{o}) \times \log \pi^\theta(a, \mathbf{o}) \tag{5}$$

3.3 データの共有割合

非類似度と未学習度を用いて，データの共有割合と優先度を決定する．エージェント i と j の共有割合 $X_{i,j}$ は，パラメータ δ と σ，未学習度 $u_{i,j}^{UL}$，非類似度 $u_{i,j}^{\mathrm{DS}}$ を用い，次式で決定する．

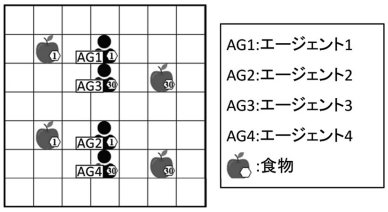

Fig. 1 環境 1 の初期位置

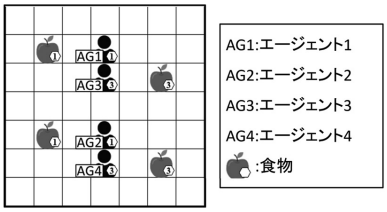

Fig. 2 環境 2 の初期位置

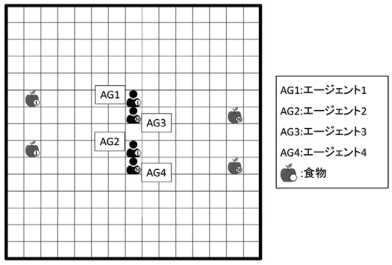

Fig. 3 環境 3 の初期位置

$$X_{i,j} = \exp\left\{-\frac{(u_{i,j}^{UL})^2}{2\delta^2}\right\}\left\{-\frac{(u_{i,j}^{\mathrm{DS}})^2}{2\sigma^2}\right\} \tag{6}$$

エージェントの行動傾向が類似し，学習が進むにつれて非類似度 $u_{i,j}^{\mathrm{DS}}$ と未学習度 $u_{i,j}^{UL}$ が 0 に近づき，データ共有割合 $X_{i,j}$ は高くなる．

3.4 共有データの選択

本論文では，累積報酬と行動傾向の二つの基準に基づいて共有データを選定する．

3.4.1 累積報酬に基づく選択

累積報酬に基づく選択では，より多くの報酬を得る行動を選択した経験を共有し，学習効率を向上させる．エージェント i のエピソード k の経験データ $Exp_{i,k}$ に対する優先度 $p^{Exp_{i,k}}$ は，次式で累積報酬として求める．

$$p^{Exp_{i,k}} = r_1^{i,k} + r_2^{i,k} + \cdots + r_T^{i,k} \tag{7}$$

3.4.2 行動傾向に基づく選択

行動傾向に基づく選択では，類似するエージェントの類似性の高い経験データを共有し，学習効率を向上させる．エージェント間の行動傾向の類似性は，特定の観測 $o_i, o_j \in \mathcal{P}_{i,j}$ を用いた非類似度計算で求められる．エージェント j のエピソード e の観測履歴 $H_j^{EP_{A,e}} \subseteq H_j^{EP_A}$，エージェント i,j の観測の集合 $H_{i,j} = \{\mathbf{o}_i, \mathbf{o}_j \,|\, (\mathbf{o}_i, \mathbf{o}_j) \in \mathcal{P}_{i,j}\}$ を用いて，エージェント j におけるエピソード e の経験データ $Exp_{j,e}$ の優先度 $p^{Exp_{j,e}}$ を次式のように定義する．

$$p^{Exp_{j,e}} = \frac{|H_j^{EP_{A,e}} \cap H_{i,j}|}{|H_j^{EP_{A,e}}|} \tag{8}$$

本論文では，(7), (8) 式のいずれかに従い，エージェント j の経験データの優先度を計算し，優先度の高い経験データから順に，$\lfloor X_{i,j} \cdot EP_A \rfloor$ 単位がエージェント i に共有される．

4. シミュレーション

4.1 実験環境

本論文では，Level-Based Foraging (LBF) 環境 [15] を用いて提案手法の有用性を示す．LBF では，エージェントが領域内の複数の食物を収穫し，食物のレベルに比例した報酬を得る．収穫は，エージェントのレベルの合計が食物のレベル以上の場合に成功し，それ以外の場合は失敗する．収穫に成功した場合，次式により報酬が決定される．

$$r_t^i = \frac{fl_n}{\sum_{k\in\mathcal{F}} fl_k} \times \frac{al_i}{\sum_{l\in\mathcal{N}^n} al_l} \tag{9}$$

ここで f_{ln} は n 番目の食物のレベル，a_{li} はエージェント i のレベル，F は食物の集合，N_n は n 番目の食物を収穫したエージェントの集合を表す．(9) 式により，エージェントは自身のレベルに比例する報酬を得る．エージェントは自身と食物の位置，食物のレベルを観測でき，移動，収穫，何もしないのいずれかを行動として選択する．また，収穫以外の行動を選択した場合や収穫に失敗した場合にはペナルティとして負の報酬が与えられる．

本実験では，初期位置とエージェントのレベルが異なる，Figs. 1〜3 に示される 3 種類の環境で実験を行う．これらの環境ではエージェントと食物のレベル差があり，エージェントの配置も異なる．高レベルのエージェントを「高レベル」，低レベルのエージェントを「低レベル」，高レベルの食物を「高報酬」，低レベルの食物を「低報酬」と呼称する．

Fig. 1 の環境はエージェントと食物のレベルが 1 と 30 で，報酬差が大きいシンプルな環境である．Fig. 2 の環境はエージェントと食物のレベルが 1 または 3 で，報酬差が小さい環境であり，報酬差の違いによる手法の有用性を検証するために使用する．Fig. 3 の環境はマップサイズが 15×15 マスと大きく，食物の収穫までの経路が長いため，探索が困難で報酬が疎な環境である．環境 1–3 に関するパラメータを Table 1 に示す．

Table 1 環境パラメータ

	環境 1	環境 2	環境 3
マップサイズ	7×7		15×15
エージェント数	4		
食物の数	4		
ステップペナルティ	0.0001	0.001	0.0001
収穫失敗ペナルティ	0.001	0.02	0.001

シミュレーションでは，データ共有を行わずに各エージェントが独立して学習するアクター・クリティック手法による「独立学習」，Hayashida *et al.*[9] の固定割合

方式，行動傾向に基づく共有データ選択，累積報酬に基づく共有データ選択の 4 手法を比較する．各手法のパラメータは「学習サイクル」,「データ共有実行サイクル」,「データ共有開始エピソード」の三つで，Hayashida *et al.*[9] の設定に従う．すなわち，割引率を $\gamma = 0.9$，Actor および Critic の学習率を 0.00001，損失関数におけるエントロピー項の調整パラメータを 0.001 とする．固定割合方式では，300 エピソードからデータ共有を開始し，データ共有割合を $X_A = 0.25$，共有される経験データのうち，上位のものを 80%，下位のものを 20% とする．提案手法では，累積報酬あるいは行動傾向に基づいて共有されるデータを選択する 2 種類の手法を検討するが，それぞれ $EP_B = 1$ とする．また，環境 1 では $\delta = 1, \sigma = 1.0 \times 10^{-5}$，環境 2 では $\delta = 0.6, \sigma = 1.0 \times 10^{-5}$，環境 3 では $\delta = 2.5, \sigma = 2.0 \times 10^{-5}$ とする．

1 エピソードの最大ステップ数を 50，3000 エピソードを 1 試行とし，各環境に対してそれぞれの手法を適用したシミュレーション実験を 10 試行行う．

4.2 シミュレーション実験

4.2.1 環境 1

環境 1 の結果を Fig. 4 に示す．縦軸は 100 エピソードごとの平均ステップ数，横軸はエピソードである．

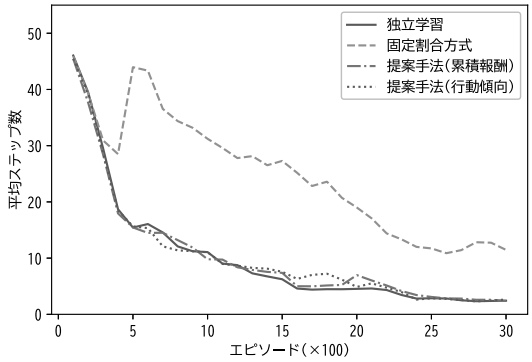

Fig. 4　環境 1 のステップ数の平均

Fig. 4 より，固定割合方式はステップ数が多く適切な行動を学習できなかったが，独立学習，提案手法（行動傾向），提案手法（累積報酬）は学習が進むとステップ数が減少した．環境 1 ではエージェント間の報酬差が大きいため，高レベルは高報酬を得やすく，低レベルは得にくい．固定割合方式では高レベルのデータを低レベルに共有するため，低レベルが適切な行動を学習できなかったと考えられる．一方，提案手法（行動傾向）と提案手法（累積報酬）は異なるレベルのエージェント間でのデータ共有が少ないため，適切に行動でき，固定割合方式より良い結果を得た．独立学習も同等の結果を示し，ステップ数は最適値に近づいたが，環境 1 が学習しやすい環境であったことが原因と考えられる．

環境 1 において，2 種類の提案手法を用いた実験の結果として，エージェント 1, 3 のデータ共有割合の推移を，横軸をエピソード，縦軸を共有割合とする Figs. 5~8 に示す．なお，共有割合について，エージェント 2 はエージェント 1 と，エージェント 4 はエージェント 3 とそれぞれ同様の結果であった．

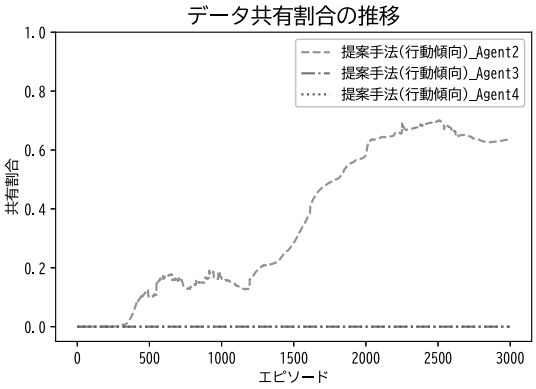

Fig. 5　共有割合（エージェント 1: 提案手法（行動傾向））

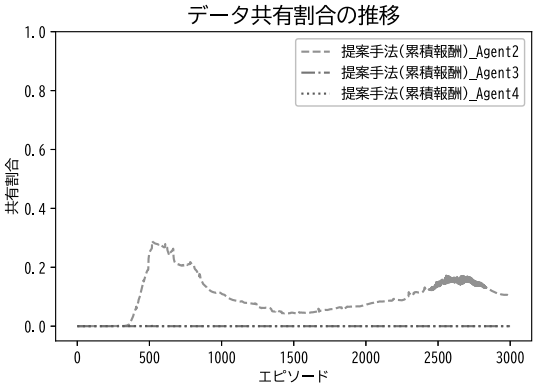

Fig. 6　共有割合（エージェント 1: 提案手法（累積報酬））

Figs. 5~8 より，提案手法（行動傾向），提案手法（累積報酬）は，特徴が異なるエージェント間ではデータ共有を抑制し，特徴の類似するエージェント間でデータ共有を適切に実施していることが確認できる．これにより，エージェントの特徴に基づく選択的なデータ共有が実現され，学習効率を向上させている．一方固定割合方式では，特徴が異なるエージェント間で不適切なデータ共有が行われ，学習効率を低下させる傾向がみられた．提案手法はこれを効果的に回避し，性能悪化を防いでいる．

4.2.2 環境 2

環境 2 の結果として縦軸を 100 エピソードごとの 1 エピソードで経過したステップ数の平均値，横軸をエピソードとしたステップ数の平均の推移を Fig. 9 に，累積報酬の推移を Figs. 10, 11 に示す．

Fig. 9 に示されるように，収束後のステップ数はすべての手法で近い値となっているが，固定割合方式ではデータ共有を開始した直後の 300 エピソード付近においてステップ数が増加し，一時的な学習悪化が確認された．また，Figs. 10, 11 より，固定割合方式では低レベルエー

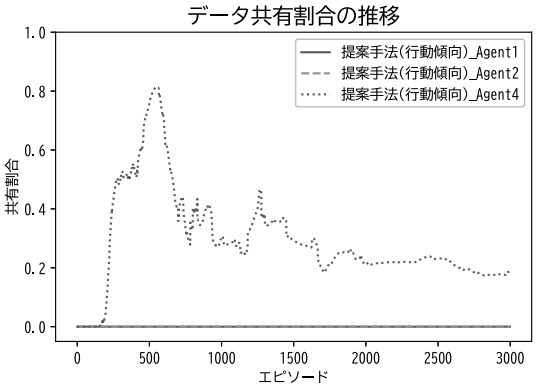

Fig. 7 共有割合（エージェント 3: 提案手法（行動傾向））

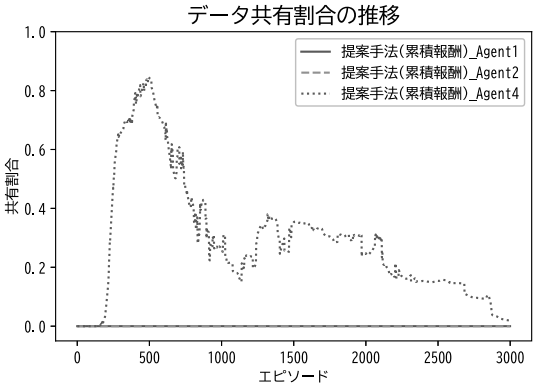

Fig. 8 共有割合（エージェント 3: 提案手法（累積報酬））

ジェントにおいて報酬値が他の手法に比べて低下していることがわかる．

環境 2 では，エージェント間の報酬差が小さいため，固定割合方式による高レベルエージェントの経験データ共有が低レベルエージェントに与える影響は限定的である．そのため，最終的な収束結果においては他の手法との差が縮小していると考えられる．一方，提案手法（行動傾向），（累積報酬）は，特徴の異なるエージェント間でのデータ共有を抑制しており，固定割合方式に比べて安定した学習を示している．平均ステップ数および獲得報酬は，提案手法と独立学習が近い結果であった．これは環境 2 が比較的学習が容易であり，データ共有がなくても安定した学習が可能であることを示している．

4.2.3 環境 3

環境 3 の実験結果として，縦軸を 100 エピソードごとの 1 エピソードで経過したステップ数の平均値，横軸をエピソードとしたステップ数の平均の推移を Fig. 12 に，縦軸を 100 エピソードごとの低レベル，高レベルの合計食物収穫数，横軸をエピソードとした食物収穫数の推移を Figs. 13, 14 に，縦軸を 100 エピソードごとの低レベル，高レベルの累積報酬値，横軸をエピソードとした累積報酬値の推移を Figs. 15, 16 に示す．

Fig. 13 に示されるように，固定割合方式ではデータ共有開始後の 300 エピソード以降に収穫数が減少してお

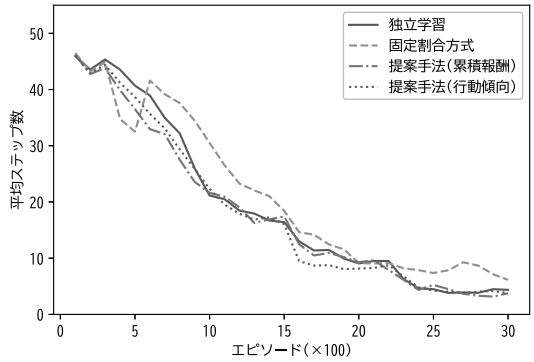

Fig. 9 環境 2 のステップ数の平均

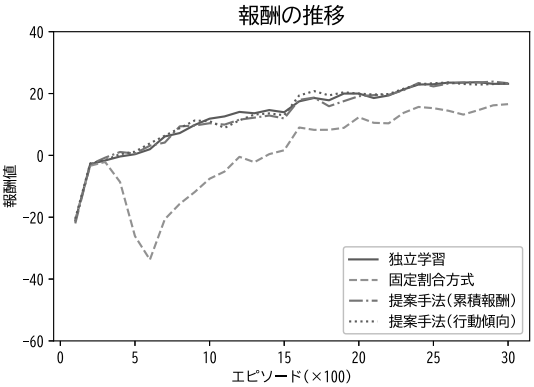

Fig. 10 低レベル

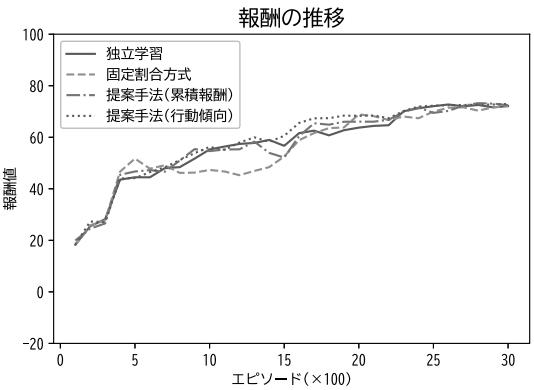

Fig. 11 高レベル

り，他の手法に比べ学習効率が著しく低下している．一方，独立学習および提案手法（行動傾向），（累積報酬）は，低レベルエージェント，高レベルエージェントの両方の収穫数が増加している．特に提案手法（累積報酬）は，低レベルエージェントにおいて最も多くの収穫数を達成している．また，高レベルエージェントにおいても，Fig. 16 に示されるように，報酬値が固定割合方式および独立学習を上回っている．

提案手法（累積報酬）は，エージェントの特徴に基づいた選択的なデータ共有に加え，累積報酬を基準としたデータ選別を行うことで，高い報酬を得る経験データを

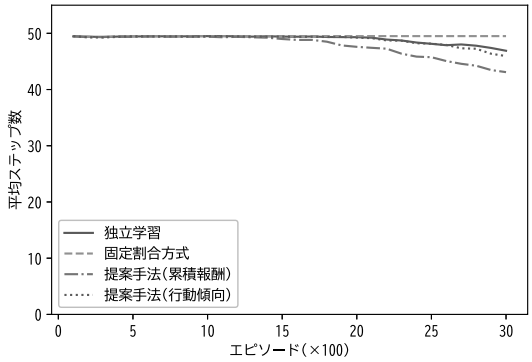

Fig. 12　環境 3 のステップ数の平均

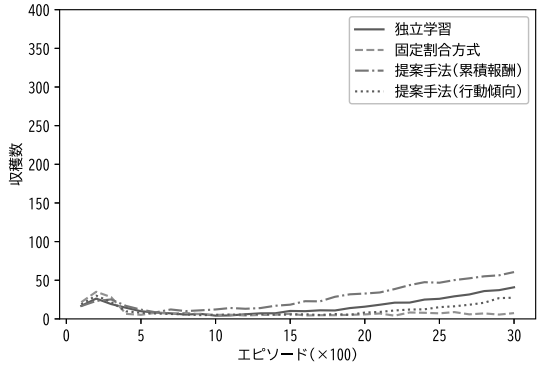

Fig. 13　低レベル

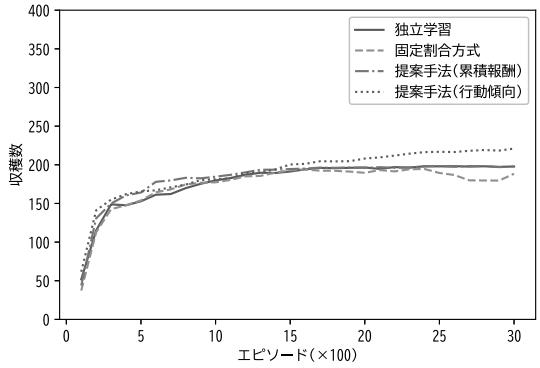

Fig. 14　高レベル

効果的に共有している．これにより，収穫経験が少ない低レベルエージェントにおいて，収穫成功までの一連の行動を強化する学習が進み，独立学習や提案手法（行動傾向）を上回る結果が得られたと考えられる．

一方，提案手法（行動傾向）は，収穫成功経験が少ない低レベルエージェントにおいて，収穫失敗の経験データが共有される割合が高かったと考えられる．このため，収穫行動を強化する学習が遅れ，収穫数が他手法よりも少ない結果となった．しかし，高レベルエージェントでは，エージェントの特徴に基づいた適切なデータ共有が行われた結果，収穫数および報酬値が独立学習や固定割

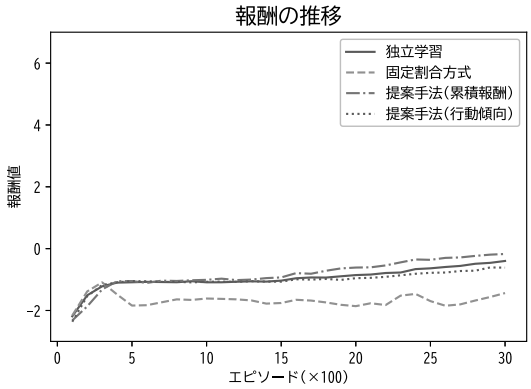

Fig. 15　低レベル

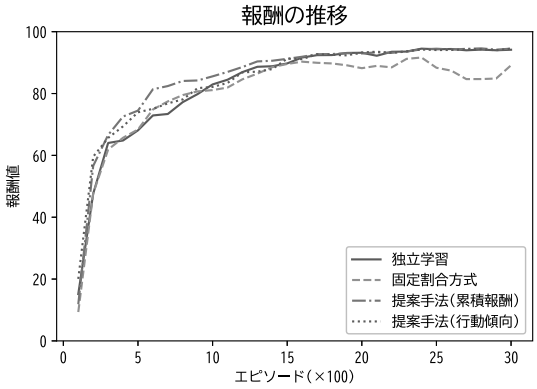

Fig. 16　高レベル

合方式を上回っている．これらの結果は，非対称性の強い環境における提案手法の学習効率向上効果を明確に示している．

ここまでに述べたシミュレーション実験 1–3 の結果から，提案手法の有用性が以下の特徴に基づくものであることが示されたといえる．提案手法は，エージェント間の観測や行動の類似性に基づいて共有されるデータを選別することで，特徴が異なるエージェント間での不適切なデータ共有を抑制し，学習効率の低下を回避することができる．また，累積報酬を基準としたデータ選別を行うことで，高い報酬を得る経験を効果的に共有し，収穫行動の学習を強化している．このように，提案手法は非対称性の強い環境において特にその有用性が顕著であり，学習効率を向上させることできる．さらに提案手法は，非対称性が低い環境においても安定した性能を維持しており，環境特性に応じた適応的な学習が可能であることが示された．これらの特徴は，提案手法が多様な条件下で安定的かつ効率的な学習を実現するための基盤技術として有用であることを示唆している．

5.　おわりに

本論文では，アクター・クリティック手法に基づき，複数のエージェント間で選択的に経験データを共有する手法を構築し，報酬に差がある LBF 環境でシミュレーショ

ン実験を行った．Hayashida *et al.*[9] の固定割合方式では，異なる特徴を持つエージェント間でデータを共有するため，学習効率の向上が難しかった．一方，提案手法（行動傾向，累積報酬に基づくデータ選択）は，エージェントの特徴に基づくデータ共有で固定割合方式より多くの食物を収穫し，高い報酬を得た．特に環境 3 では，提案手法が独立学習や固定割合方式よりも優れた成果を示した．

未学習度は，学習が容易な環境で共有率の調整に貢献したが，学習が困難な環境では方策の更新が遅く，データ共有の調整が難しかった．今後は，未学習度の算出に方策分布のエントロピー変化率や TD 誤差の活用を検討する．

参 考 文 献

[1] S. S. Shwartz, S. Shammah and A. Shashua: Safe, multi-agent, reinforcement learning for autonomous driving, *ArXiv*, 1610.03295v1 (2016)

[2] Y. Ye, D. Qui, M. Sun, D. Papadaskalopoulos and G. Strbac: Deep reinforcement learning for strategic bidding in electricity markets; *IEEE Transactions on Smart Grid*, Vol. 11, pp. 1343–1355 (2020)

[3] E. Namalomba, H. Feihu and H. Shi: Agent based simulation of centralized electricity transaction market using bi-level and Q-learning algorithm approach; *International Journal of Electrical Power & Energy Systems*, Vol. 134 (2021)

[4] J. Orr and A. Dutta: Multi-agent deep reinforcement learning for multi-robot systems: A survey; *Sensors*, Vol. 23 (2023)

[5] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel and I. Mordatch: Multi-agent actor-critic for mixed cooperative-competitive environments; *Proceedings of the 31st Conference on Neural Information Processing Systems*, Vol. 30, pp. 6382–6393 (2017)

[6] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. A. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg and D. Hassabis: Human-level control through deep reinforcement learning; *Nature*, Vol. 518, pp. 529–533 (2015)

[7] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver and D. Wierstra: Continuous control with deep reinforcement learning, *arXiv*, 1509.02971 (2015)

[8] T. Rashid, M. Samvelyan, C. S. D. Witt, G. Farquhar, J. N. Foerster and S. Whiteson: QMIX: Monotonic value function factorisation for deep multiagent reinforcement learning; *Proceedings of the 35th International Conference on Machine Learning*, Vol. 80, pp. 4295–4304 (2018)

[9] T. Hayashida, I. Nishizaki, S. Sekizaki and Q. Liu: Adaptive data sharing methods for multi-agent systems using deep reinforcement learning; *International Journal of Computational Intelligence Studies (Special Issue on IEEE IWCIA 2021 Innovative Methods and Applications in Computational Intelligence)*, Vol. 11, pp. 176–199 (2022)

[10] M. Gerstgrasser, T. Danino and S. Keren: Selectively sharing experiences improves multi-agent reinforcement learning; *Proceedings of the 22nd International Conference on Autonomous Agents and Multiagent Systems*, pp. 2433–2435 (2023)

[11] 伊藤: 現場で使える! Python 深層強化学習入門: 強化学習と深層学習による探索と制御, 翔泳社 (2019)

[12] 伊庭: 進化計算と深層学習-創発する知能, オーム社 (2015)

[13] M. Tan: Multi-agent reinforcement learning: Independent vs. cooperative agents; *Proceedings of the 10th International Conference on Machine Learning*, pp. 330–337 (1993)

[14] L. Matignon, G. J. Laurent and N. Le Fort-Piat: Independent reinforcement learners in cooperative Markov games: a survey regarding coordination problems; *The Knowledge Engineering Review*, Vol. 27, pp. 1–31 (2012)

[15] P. Georgios, C. Filippos, L. Schäfer and S. V. Albrecht: Benchmarking multiagent deep reinforcement learning algorithms in cooperative tasks; *Proceedings of the 35th Neural Information Processing Systems, Datasets and Benchmarks Track*, Vol. 34, pp. 8753–8765, (2021)

著 者 略 歴

林田 智弘 (正会員)

2009 年広島大学大学院工学研究科（複雑システム工学専攻）博士課程後期修了．広島大学大学院工学研究科（現在，先進理工系科学研究科）助手，助教，准教授を経て，2024 年同大学教授となり現在に至る．おもに，ニューラルネットワークや強化学習などを用いた，意思決定やゲーム理論などを基礎としたマルチエージェントシステムに関する研究に従事．博士（工学）．電気学会，日本オペレーションズ・リサーチ学会，IEEE などに所属．

浅野 光太郎

2022 年広島大学工学部卒業，同大学大学院先進理工系科学研究科入学，現在に至る．強化学習およびマルチエージェントシミュレーションに関する研究に従事．修士（工学）．

関崎　眞也 (正会員)

2013 年名古屋工業大学大学院工学研究科創成シミュレーション工学専攻博士後期課程修了．広島大学大学院工学研究科（現在，先進理工系科学研究科）助教を経て，2024 年同大学准教授となり，現在に至る．主として電力市場における経済主体の意思決定，送配電系統における電力品質管理に関する研究に従事．博士（工学）．電気学会，日本オペレーションズ・リサーチ学会，IEEE などに所属．

西崎　一郎

1984 年神戸大学大学院工学研究科修士課程（システム工学専攻）修了．同年，新日本製鐵株式会社入社，京都大学助手，摂南大学助教授，広島大学助教授，教授を経て，2024 年同大学名誉教授となり現在に至る．ゲーム理論および意思決定に関する研究に従事．博士（工学）．