# 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

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

**尾崎 悠毘（Haruhi Ozaki）**
**u320013@st.pu-toyama.ac.jp**

**富山県立大学 情報システム工学科**

December 5, 2025

# 背景と課題





## 強化学習と MAS の役割

- **背景**: 強化学習（DRL）は, 自律走行やロボット制御など様々な分野で応用が進展.
- 複数のエージェントが協調・競合するマルチエージェントシステム（MAS）の枠組みが有効.

## 課題：学習の不安定化（非定常性）

- 他エージェントの行動が不規則な変動を生み, エージェント視点での環境の不確実性が増大する.
- → この不確実性が, 個々のエージェントの学習を不安定化させ, 効率を低下させる.

# 先行研究：データ共有の課題





## データ共有の現状

- 経験データ（Data Sharing）を有効活用し，学習効率を改善する手法が研究されている．
- 従来は，全データ共有または一部共有の事例が報告されている

## 課題と着眼点

- 課題：無作為に共有するとノイズとなり，学習が不安定化するリスクがある．
- 本論文の着眼点： 共有データの選定基準として，エージェント間の類似性に着目し，質の高いデータを重視する．

# 提案手法 (I): 類似性に基づく選択





- エージェント間の**類似性**を基準とし, 共有する経験データを選択する手法を提案.
- 類似性の評価指標として, **非類似度** $u_{i,j}^{DS}$ と**未学習度** $u_{i,j}^{UL}$ を導入する.

## 1. 方策分布の非類似度 $u_{i,j}^{DS}$

**目的**: 観測 $o$ におけるエージェント $i$ と $j$ の行動傾向の差を評価.

$$u_{i,j}^{DS} = \mathbb{E}_o[tv(\pi_i(\cdot|o), \pi_j(\cdot|o))]$$

- $tv(\cdot)$: 全変動距離. 方策分布 $\pi_i$ と $\pi_j$ の距離を示す.
- **非類似度が高い** $\rightarrow$ 行動傾向が異なり, データ共有の**ノイズ**となる可能性が高い.

# 提案手法 (II): 共有データ選択のメカニズム





## 2. 未学習度 $u_{i,j}^{UL}$ とデータ共有割合 $X_{i,j}$

**未学習度 $u_{i,j}^{UL}$ の目的**: 学習初期段階の**不適切なデータ共有を抑制**する.

- 方策分布の**エントロピー**（行動の多様性）の積に基づいて計算される.
- **未学習度が高い** $\rightarrow$ 学習が不安定 $\rightarrow$ データ共有を**抑制**.

**データ共有割合 $X_{i,j}$** (提案手法の核心)

$$X_{i,j} = u_{i,j}^{DS} \cdot u_{i,j}^{UL}$$

- $X_{i,j}$ が低いほど（類似性が高く、学習が進んでいる）、そのエージェントの経験を**共有しない** (抑制).

# 提案手法 (III): 共有データ選択基準の確認





本研究では, 提案手法の有効性を検証するため, 以下の 2 つの基準を用いて共有データを決定しました.

## 共有データ選択の基準（比較手法）

- **累積報酬に基づく選択 (提案手法)**:
  - エピソードの**累積報酬が高い**経験を優先的に共有.
  - 高レベルエージェントの成功体験を、低レベルエージェントに効率良く共有する狙い.
- **行動傾向に基づく選択 (提案手法)**:
  - 類似性の高い観測を含む経験を優先的に共有.
  - 非類似性の高い経験データ（ノイズの可能性が高いデータ）を抑制する目的.

## 次セクションの比較手法

**独立学習**, **固定割合方式**, および上記 2 つの**提案手法**を比較します.

# 実験設定





## 実験環境

- **環境**: Level-Based Foraging (LBF) 環境を使用.
- LBF では、エージェントのレベルに応じた食物収穫タスクを行い, 報酬を得る.
- **検証環境**: 初期位置やエージェントのレベルが異なる 3 種類の非対称な環境で実験を実施.

## 比較手法

- 独立学習（データ共有なし）
- 固定割合方式（従来手法）
- 提案手法（累積報酬に基づく選択）
- 提案手法（行動傾向に基づく選択）

# 実験結果（II）：データ選択基準の比較と考察





## 累積報酬に基づく選択の優位性

- 環境 3 では，累積報酬に基づく提案手法が最も優れた結果を示した．
- 高い報酬を得る経験を効果的に共有し，収穫行動の学習が強化されたためである．

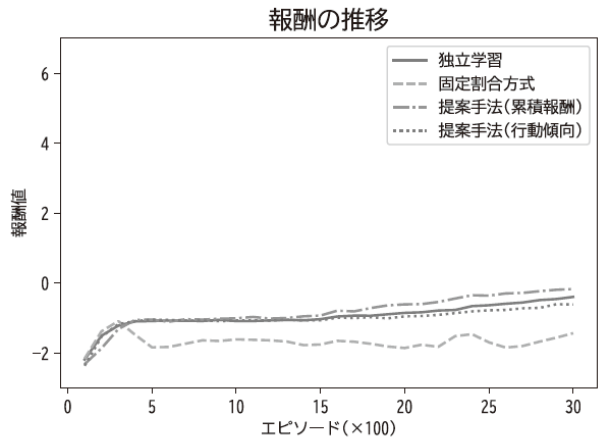


Fig. 15　低レベル

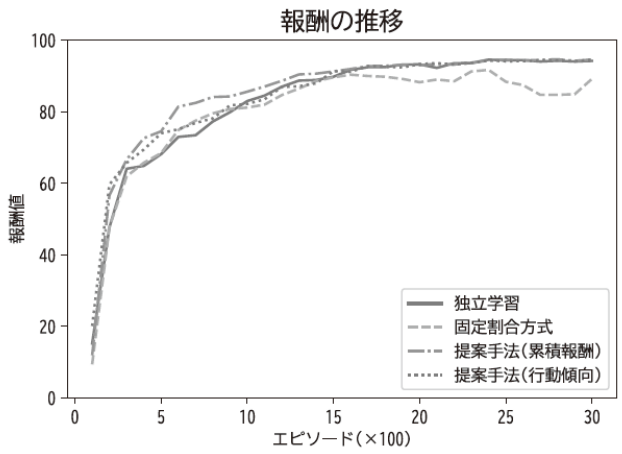


Fig. 16　高レベル

# 実験結果（II）：データ選択基準の比較と考察





## 考察

- 提案手法は, エージェント間の類似性に基づいて共有データを選別することで, 不適切なデータ共有を抑制する.
- 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, 非対称性が低い環境でも安定した性能を維持している.

# まとめ





## 本論文の主な貢献

- マルチエージェント強化学習において, 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.
- 選択的共有により, 固定割合方式よりも学習効率の向上とノイズの抑制が可能であることを実験で検証
- 非対称性の強い環境での検証で, 提案手法の有効性を明確に示した.

# 今後の課題





## 今後の課題

- 汎用性の高い類似性指標の検討： 状態表現など, より汎用的な類似性指標を用いたデータ選択手法を検討すること.
- 複雑環境での有効性の検証： 競合や部分観測を含む, より複雑で現実的なマルチエージェント環境での有効性を検証すること.
- 動的な共有機構の導入： 学習進度に応じて, 共有データの割合や重み付けを動的に調整する手法を検討すること.
- 未学習度のより高度な定量化： データ共有の判断に用いる未学習度 $u_{i,j}^{UL}$ の算出方法として, 方策分布のエントロピー変化率や TD 誤差の利用を検討すること.