二個体間協調に基づく重みづけ行動評価による マルチエージェント逆強化学習

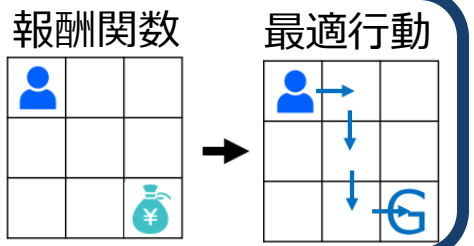
2022 年度 卒研発表

I 類 メディア情報学プログラム

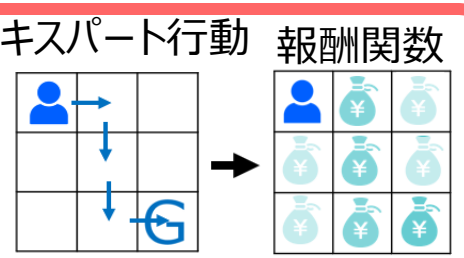
高玉研究室 1910094 植木駿介

はじめに

強化学習 [Watkins+, 1992]
- 報酬関数から行動の獲得
- **報酬設計が困難**



逆強化学習 [Ng+, 2000]
- エキスパート行動
  から報酬関数を推定



マルチエージェントシステム(MAS)における逆強化学習の問題点

- **最適・準最適なエキスパート行動を事前に用意する必要がある**

  ⇒ ナッシュ均衡解を計算して最適解を獲得

  ⇒ 環境の複雑化・エージェント数の増加に伴い，計算が困難

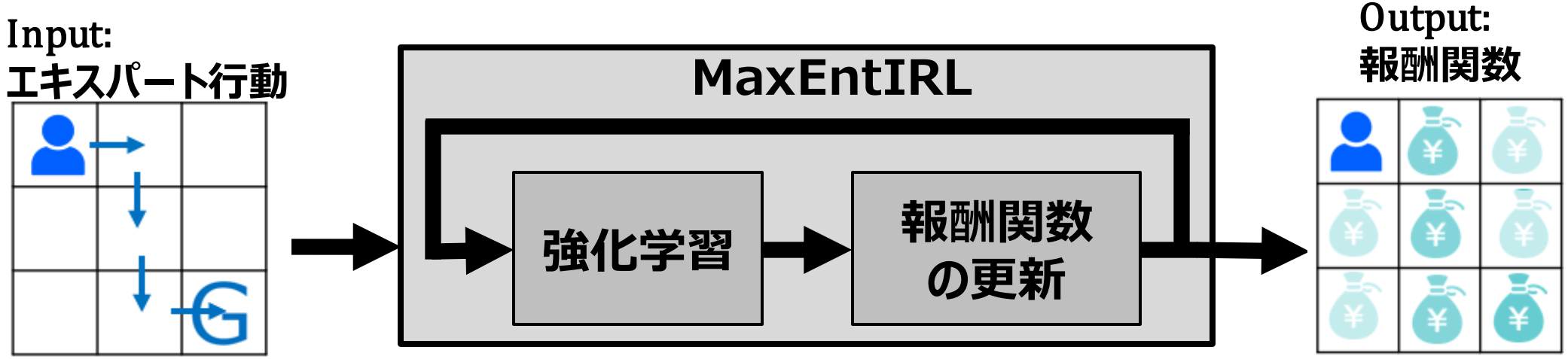
Q. 最適・準最適な行動を事前に用意するできない場合にどうするか？

本研究の目的

- **非最適エキスパート行動から、最適な報酬関数を獲得する機構を考案し，その有効性を検証する**

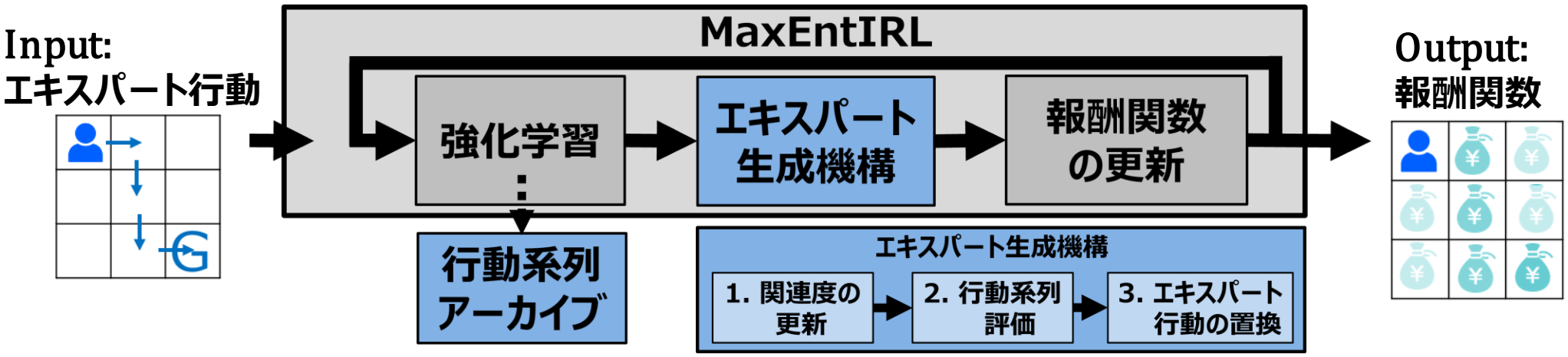
逆強化学習（従来）

- **MaxEntIRL** (Maximum Entropy IRL，[Ziebart+, 2008])
  - **強化学習により報酬関数から行動規則を獲得**
  - **以下の観点から，報酬関数を更新**
    - 行動規則によりエキスパートが到達した状態に導く確率を最大化
    - エキスパート行動が訪れていない状態は一様な確率を与える

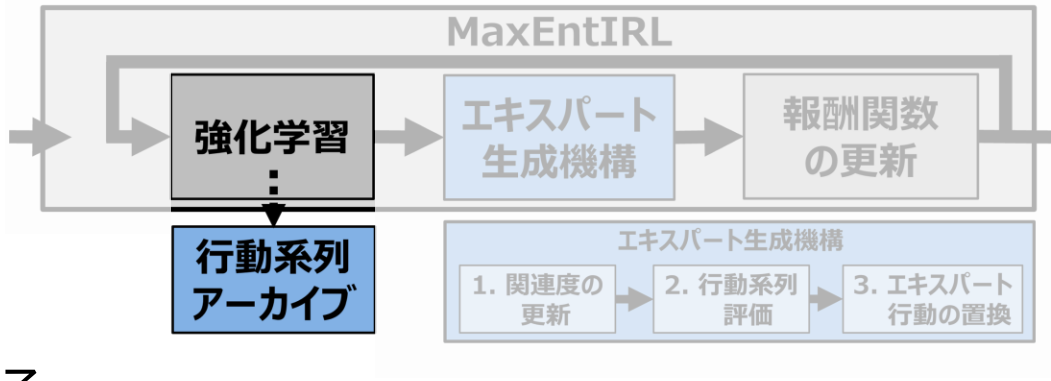


提案手法

- **WTC-MAIRL**（Weighted Two-individuals Cooperative – MAIRL）
  - **行動系列MAIRLアーカイブ**
    - 有用な行動系列の獲得
  - **エキスパート生成機構**
    - 二個体間の関連度を計算
    - 関連度が高いエージェントほど，そのエージェントと協調する行動系列を高く評価
    - 評価が最大の行動系列でエキスパート行動を置換



提案手法　WTC-MAIRL
(Weighted Two-individual Cooperative – MAIRL)



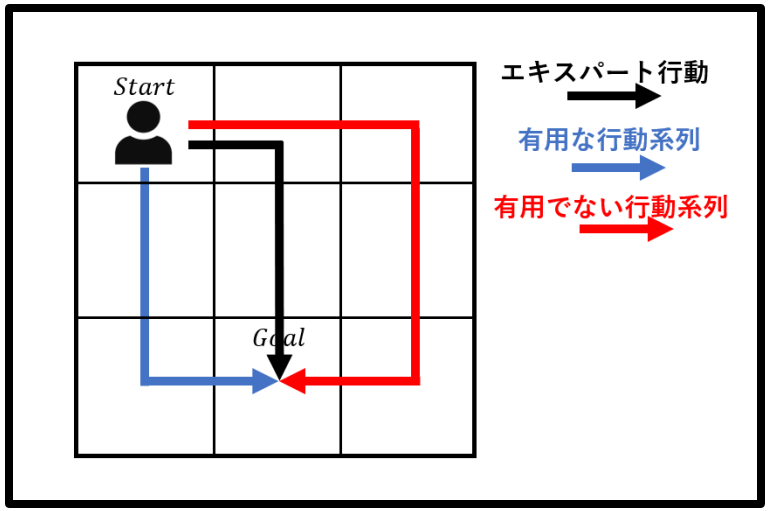
● 行動系列のアーカイブ

1．強化学習の際に，**有用な行動系列をアーカイブ**する

（有用な行動系列：エキスパート行動のステップ数以下の行動系列）

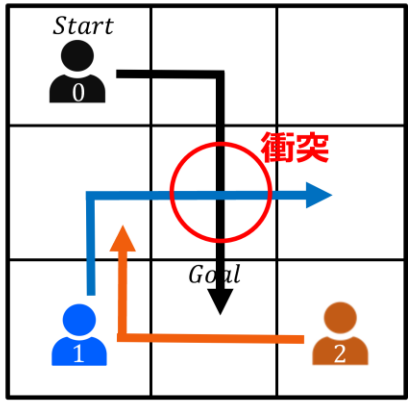
2. 行動系列を実行した際の各エージェントとの**衝突回数・非衝突回数を更新**する

迷路問題を想定した有用な行動



アーカイブの情報の更新の例

行動系列1を実行（黒）



$Agent_0$のアーカイブの例

行動系列1

各エージェントとの衝突情報

エージェント	衝突回数	非衝突回数
$Agent_1$	80+1	20
$Agent_2$	10	90+1

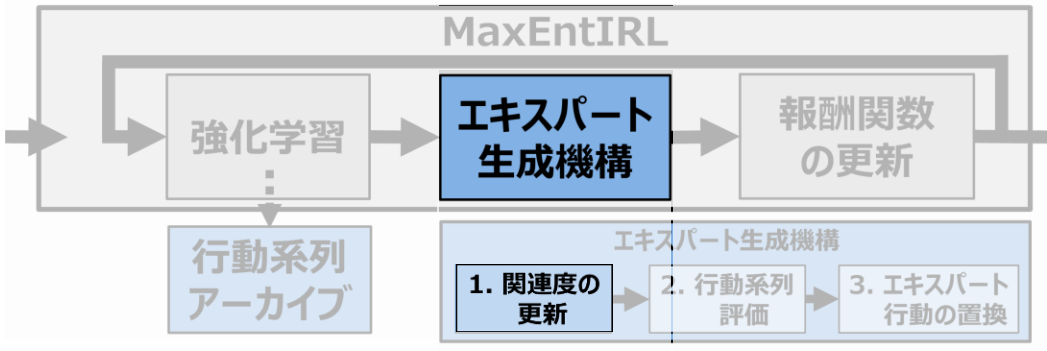
・・・

行動系列N_{traj}

各エージェントとの衝突情報

エージェント	衝突回数	非衝突回数
$Agent_1$	10	520
$Agent_2$	20	510

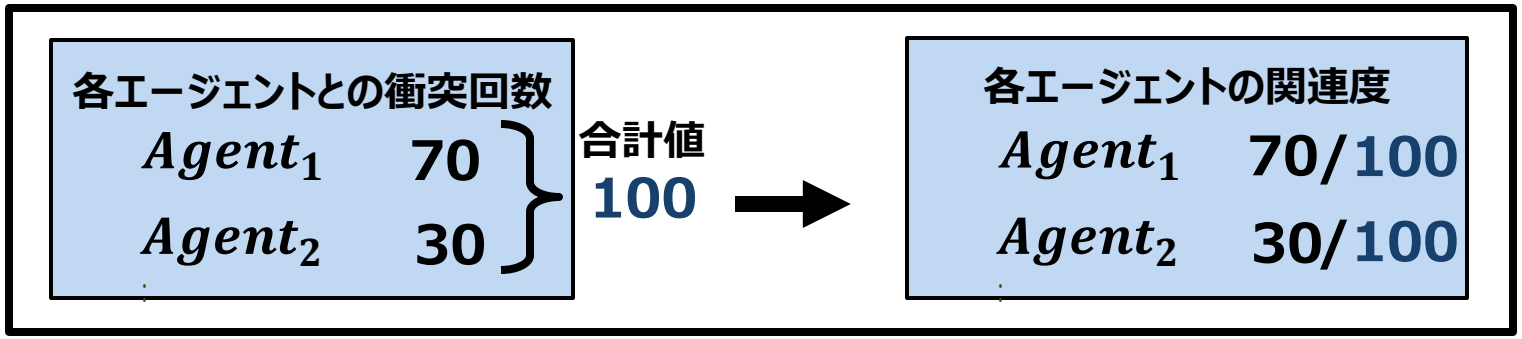
提案手法　WTC-MAIRL
(Weighted Two-individual Cooperative – MAIRL)



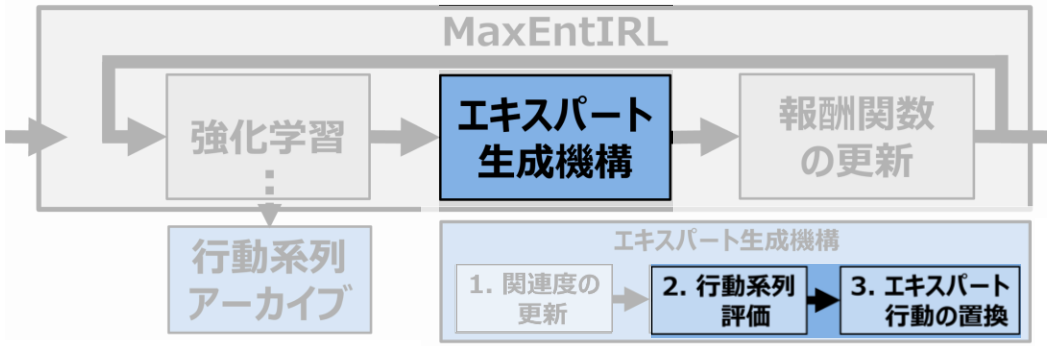
● 関連度の更新

- 関連度は**各エージェントの衝突度合い**とする
- 強化学習により得られた方策に従って行動した際の各エージェントの衝突回数を加算

$Agent_0$の各エージェントの関連度の計算



提案手法　WTC-MAIRL
(Weighted Two-individual Cooperative – MAIRL)



● 行動系列の評価
- アーカイブした行動系列の評価をする

評価式：$Eval(\zeta_k^i) = \sum_{j=0, j\neq i}^{N_{agent}-1} \underbrace{c_{agent}^{i,j}}_{\text{Weight}} \times \underbrace{ncol_{\zeta_k^i}^{i,j}}_{\text{非衝突率}}$

（＝関連度）

ζ_k^i ：$Agent_i$がアーカイブしたk番目の行動系列
$c_{agent}^{i,j}$ ：$Agent_i$の$Agent_j$との関連度
$ncol_{\zeta_k^i}^{i,j}$ ：行動系列ζ_k^iのエージェントjの非衝突率

評価値の計算例

$Agent_0$のk番目のアーカイブ

行動系列k
各エージェントとの衝突情報

	衝突回数	非衝突回数	各エージェントとの非衝突率
$Agent_1$	20	80	→ 0.8
$Agent_2$	80	20	→ 0.2

Weight(関連度)

$$Eval\left(\zeta_k^0\right) = \underbrace{0.7}_{\text{Weight}} \times \underbrace{0.8}_{\text{非衝突率}} + 0.3 \times 0.2$$
（0.7：$Agent_1$、0.3：$Agent_2$）

● エキスパート行動の置換
- 評価値が最大の行動系列をエキスパート行動と置換する

実験

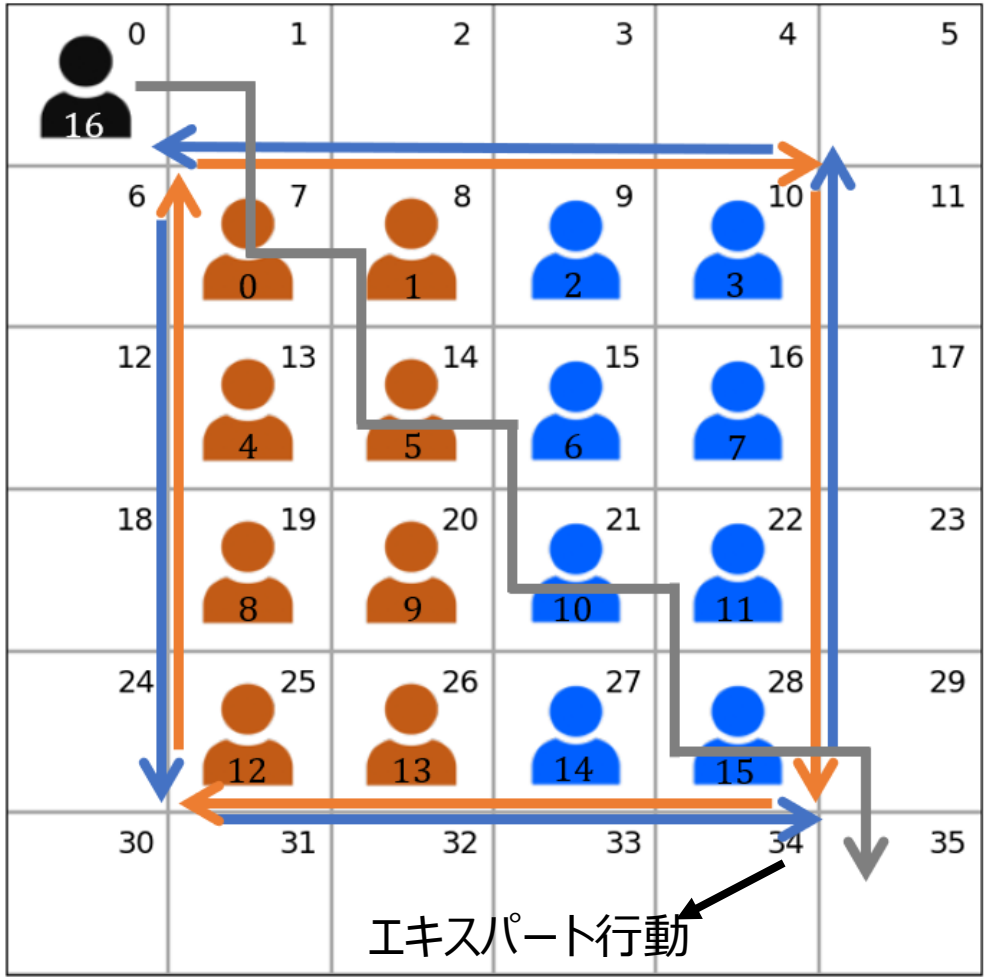
- **実験内容：従来手法と提案手法の比較**
  - 従来手法　：MaxEntIRL
  - 提案手法１：　TC-MAIRL (Unweighted)
  - 提案手法２：　WTC-MAIRL

重みづけしない評価

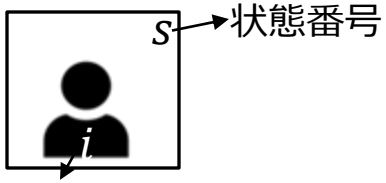
- **評価項目：**
  - 全エージェントの平均ステップ数
  - 獲得したエキスパート行動・報酬関数

- **実験設定：**
  - 対称のマスにあるゴールを設定
  - 行動は上下左右の4通り

迷路問題



橙：時計回り
青：反時計回り



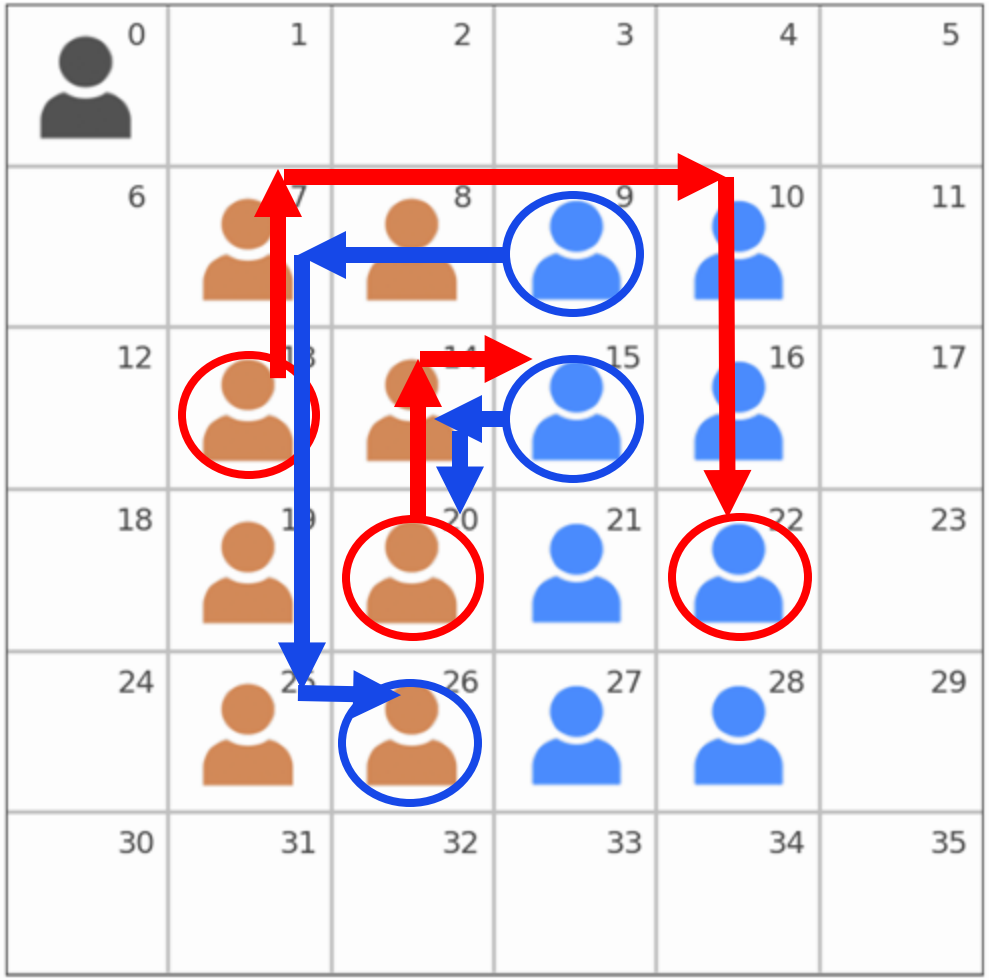
実験

- **実験内容：従来手法と提案手法の比較**
  - 従来手法　：MaxEntIRL
  - 提案手法１：　TC-MAIRL (Unweighted)
  - 提案手法２：　WTC-MAIRL

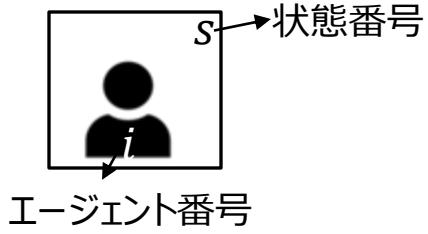
重みづけしない評価

- **評価項目：**
  - 全エージェントの平均ステップ数
  - 獲得したエキスパート行動・報酬関数

- **実験設定：**
  - 対称のマスにあるゴールを設定
  - 行動は上下左右の4通り



橙：時計回り
青：反時計回り

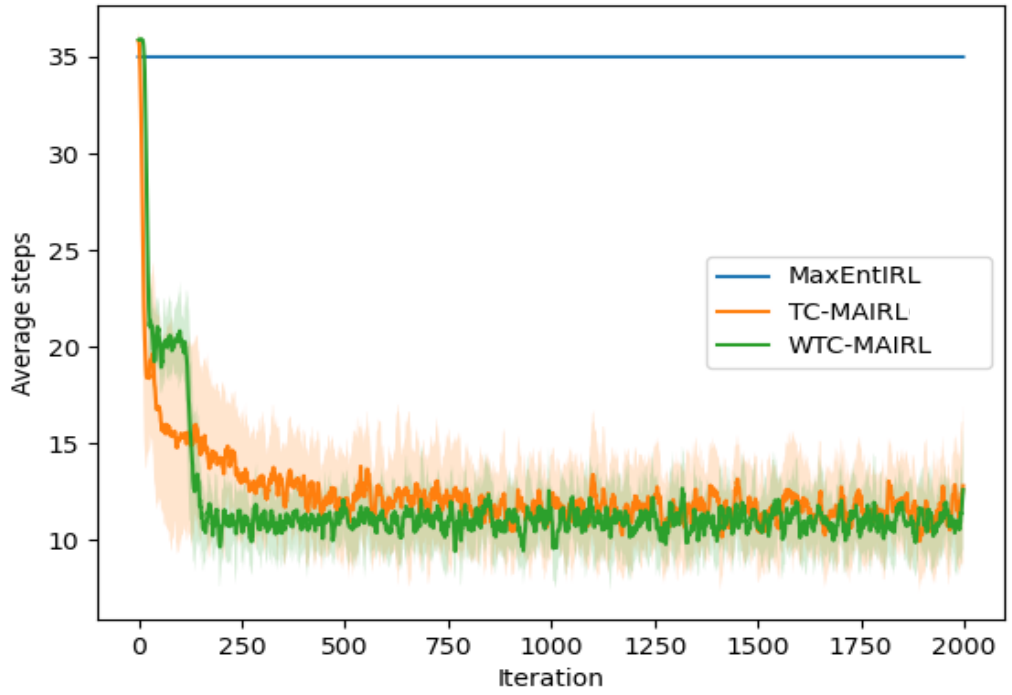


実験結果

従来手法：解決不可

提案手法：重みづけ評価により，収束速度の向上，学習の安定性が向上

平均ステップ数



標準偏差の平均値

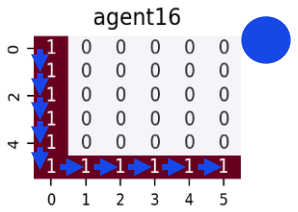
	標準偏差
TC-MAIRL	2.79
WTC-MAIRL	**1.87**

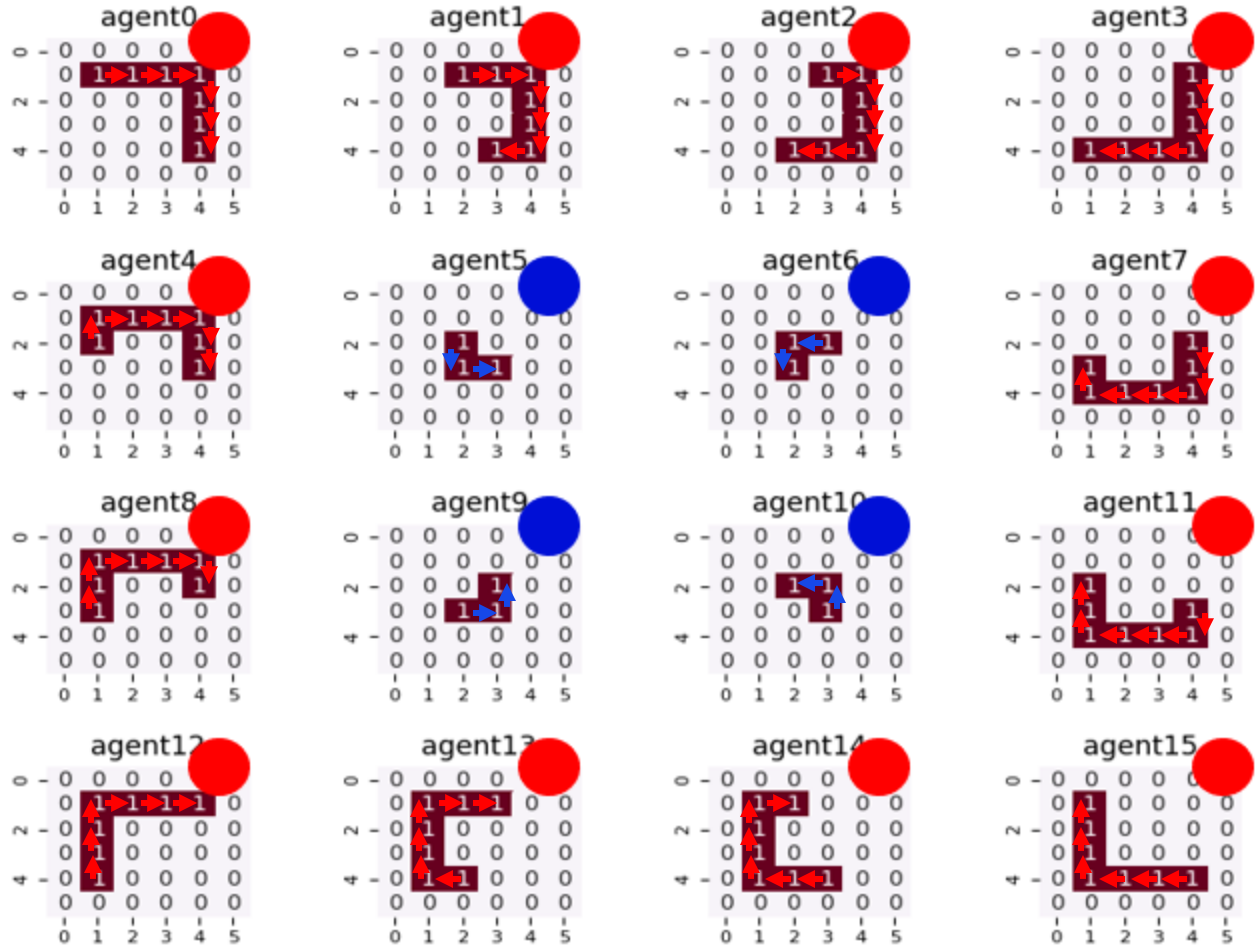
実験結果

- 最適なエキスパート行動の獲得に成功

➡ 最適なエキスパート行動の置換が可能

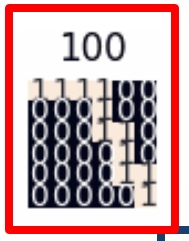
エキスパート行動の結果（ WTC-MAIRL ）





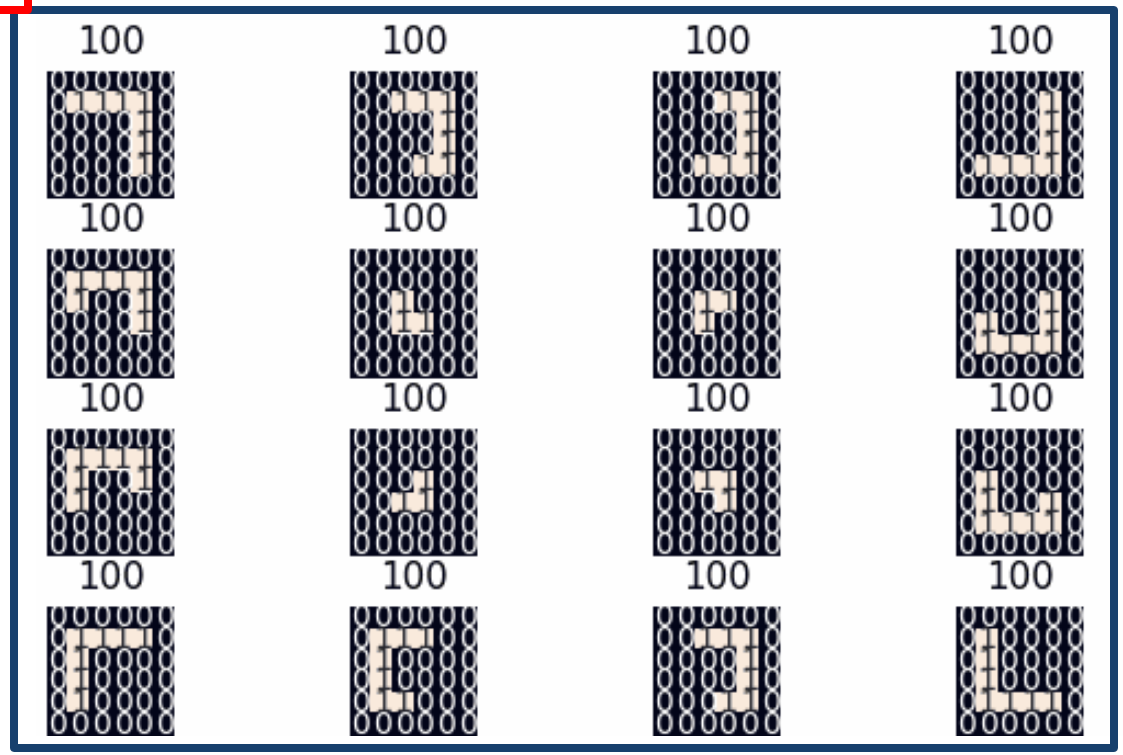
実験結果

100-200itearationエキスパート行動

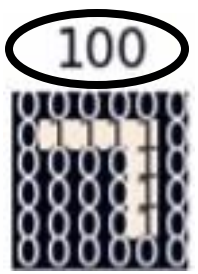


130iterationで
最適解に収束 (Agent16)

すでに最適解に収束

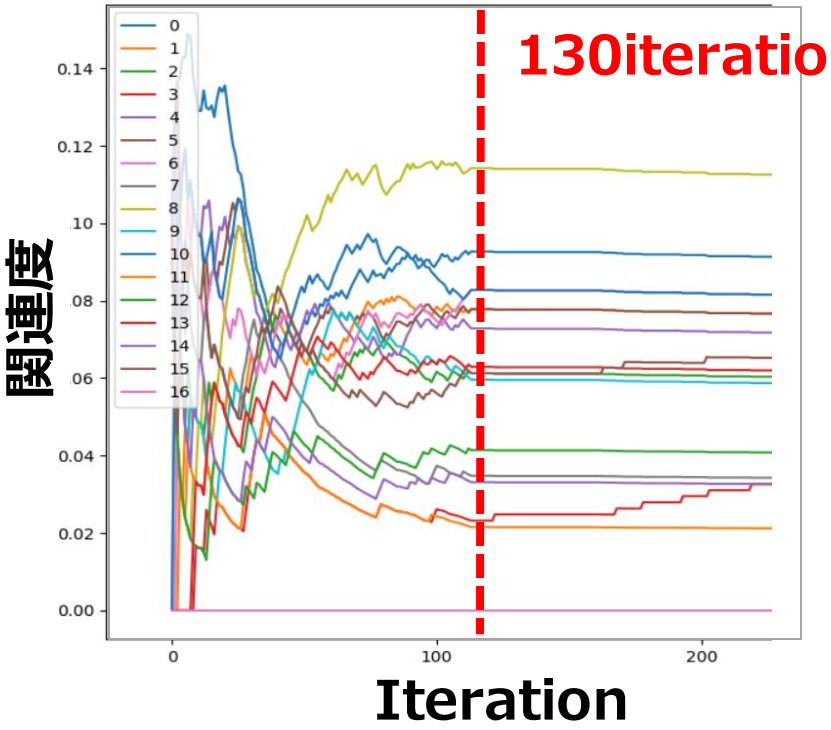


Iteration



Expert

Agent16の関連度



- **関連度による重みづけ評価により，最適解に収束可能**

おわりに

- **目的：**
  - 非最適なエキスパート行動から協調に必要な行動を導く報酬関数の獲得
- **提案：**
  - WTC-MAIRL（二個体間協調＋重みづけ評価に基づくIRL）
- **結果：**
  - 非最適なエキスパート行動から**最適なエキスパート行動の獲得に成功**
- **今後の課題：**
  - 連続空間への拡張

補足

● 行動系列の評価

- アーカイブした行動系列の評価をする

TC-MAIRLの評価式：$Eval(\zeta_k^i) = \sum_{j=0, j\neq i}^{N_{agent}-1} \underbrace{ncol_{\zeta_k^i}^{i,j}}_{\text{非衝突率}}$

WTC-MAIRLの評価式：$Eval(\zeta_k^i) = \sum_{j=0, j\neq i}^{N_{agent}-1} \underbrace{c_{agent}^{i,j}}_{\text{Weight}} \times \underbrace{ncol_{\zeta_k^i}^{i,j}}_{\text{非衝突率}}$

ζ_k^i ：$Agent_i$がアーカイブしたk番目の行動系列
$c_{agent}^{i,j}$：$Agent_i$の$Agent_j$との関連度
$ncol_{\zeta_k^i}^{i,j}$：行動系列ζ_k^iのエージェントjの非衝突率

評価値の計算例

$Agent_i$のk番目のアーカイブ

行動系列k

各エージェントとの衝突情報

	衝突回数	非衝突回数	各エージェントとの非衝突率
$Agent_1$	20	80	0.8
⋮	⋮	⋮	⋮
$Agent_N$	80	20	0.2

TC-MAIRL

$Eval(\zeta_k^i) = 0.8 + \cdots + 0.2$

WTC-MAIRL

$Eval(\zeta_k^i) = 0.8 \times c_{agent}^{i,1} + \cdots + 0.2 \times c_{agent}^{i,N}$

● エキスパート行動の置換

- 評価値が最大の行動系列をエキスパート行動と置換する

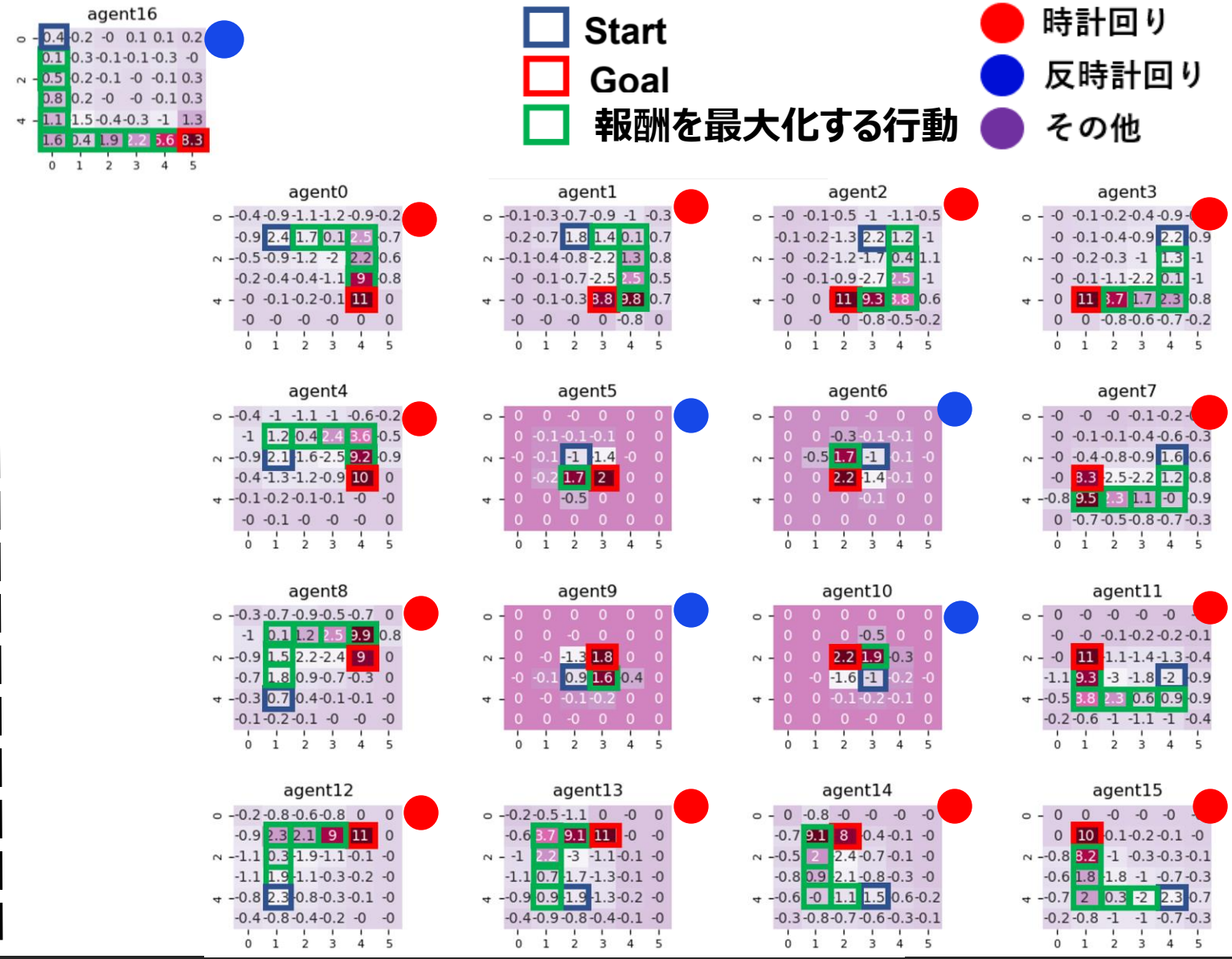
補足

- **報酬を最大化する行動**

＝ 最適な行動

➡ **最適な報酬関数の獲得に成功**

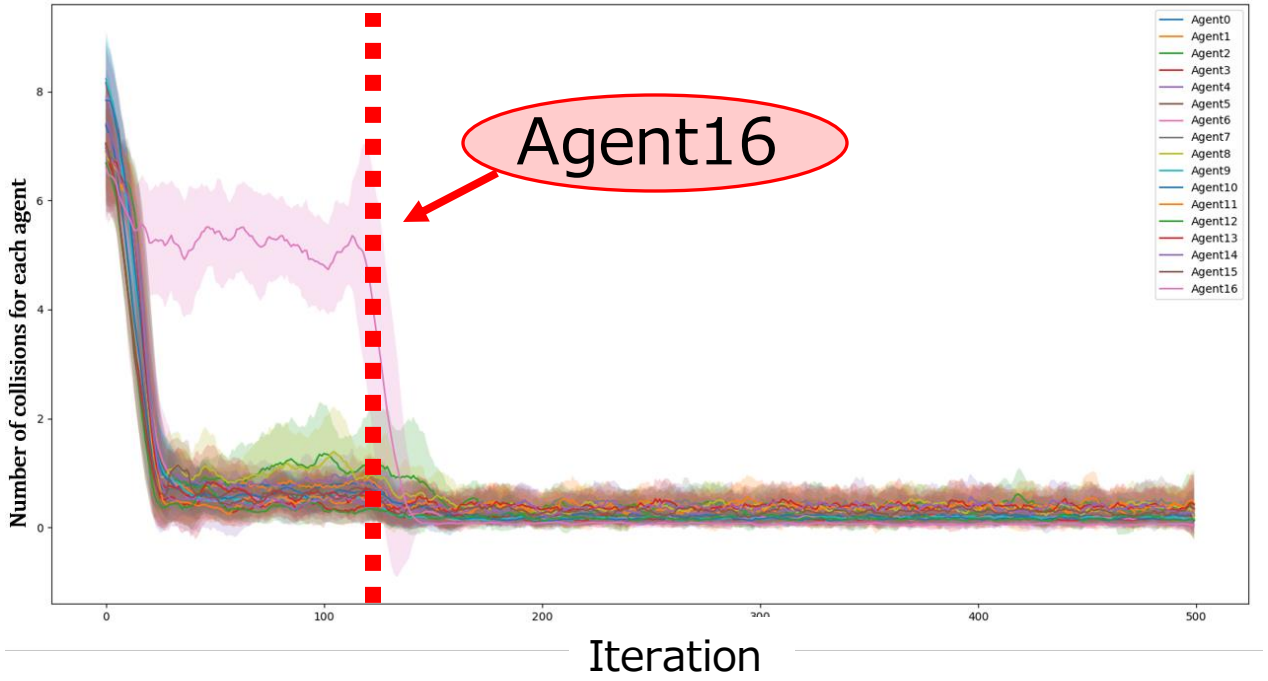
報酬関数の結果（ WTC-MAIRL ）



補足

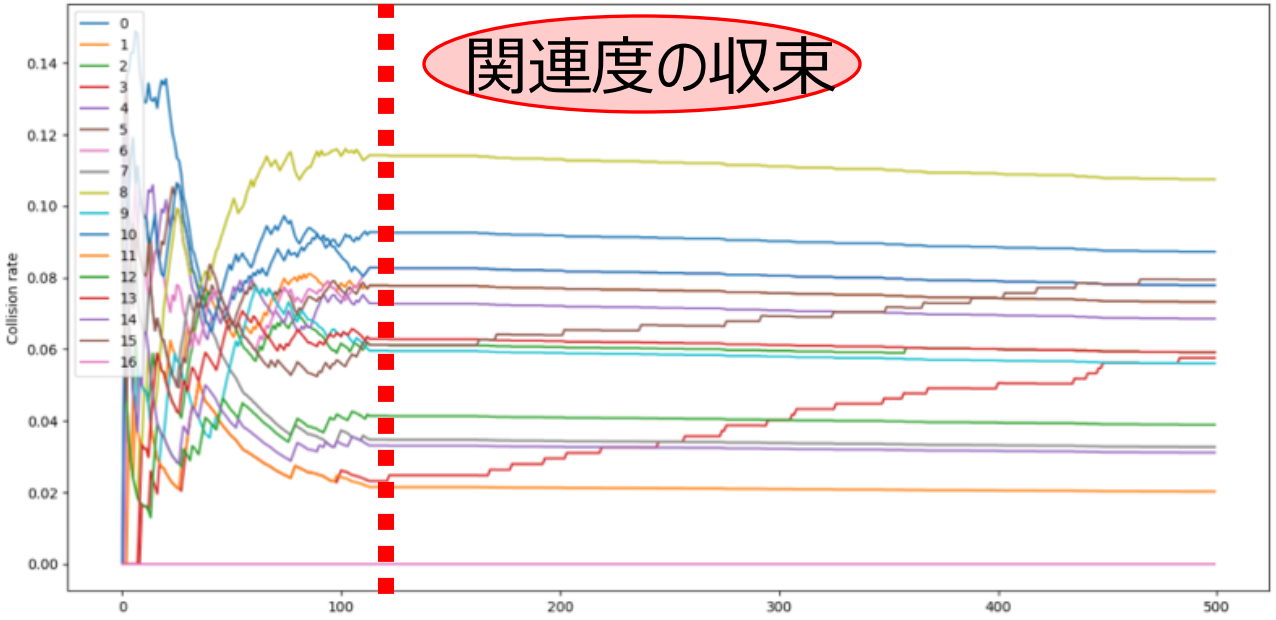
ステップ数が停滞する原因

各エージェントの衝突回数



Iteration

Agent16の関連度の変化



Iteration