Cypher研究会
（2022年2月22日）

マルチエージェント強化学習における報酬による協調制御とその展望

○上野　史†

† 
岡山大学大学院 自然科学研究科
GRADUATE SCHOOL OF NATURAL SCIENCE AND TECHNOLOGY, OKAYAMA UNIVERSITY

自己紹介

- 氏名：上野 史（自然科学学域 助教）
- 学位：博士（工学）（電気通信大学，2020年3月）
- 専門：マルチエージェント強化学習，他

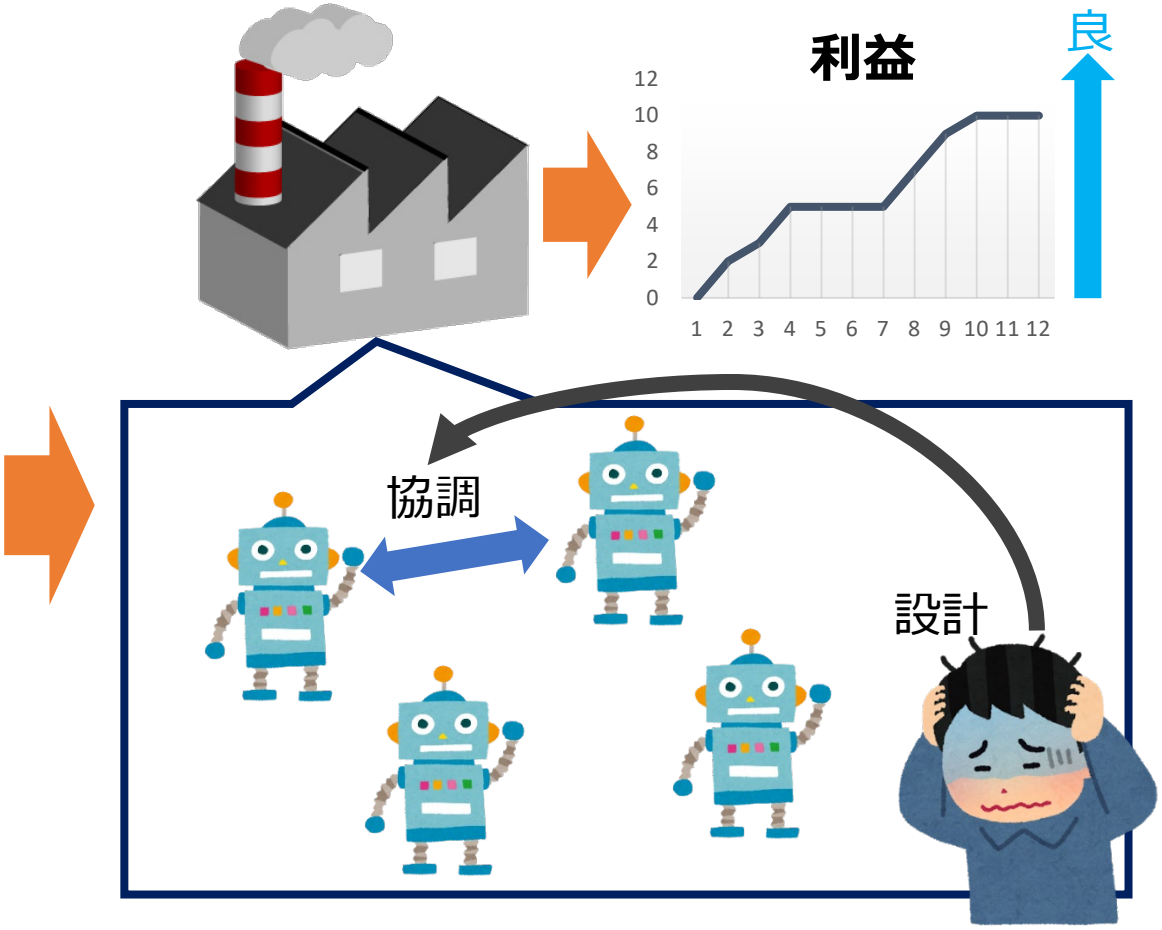
マルチエージェントシステム
(Multi-Agent System: MAS)

- 複数の活動主体（エージェント）を適切に動かすシステム
- エージェント：モデル内の活動主体（ロボットや車など）

活用例

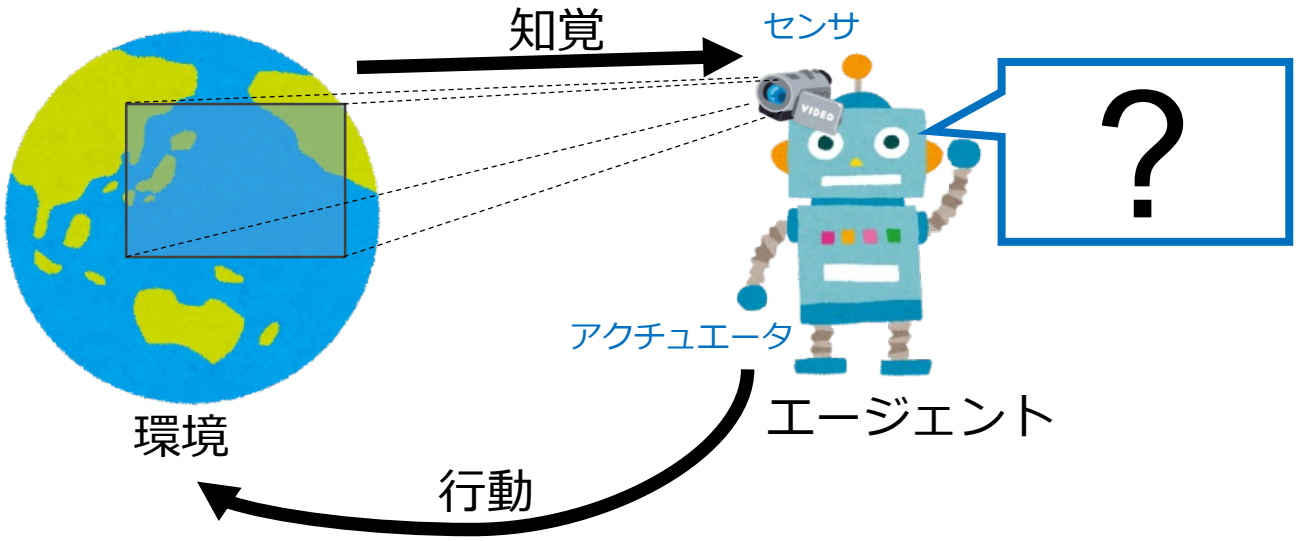
渋滞解消

ロボット制御



エージェント

- センサとアクチュエータにより環境と相互作用
  - センサによる認知
  - アクチュエータによる行動
- 知的活動：センサ情報から適切にアクチュエータを動作



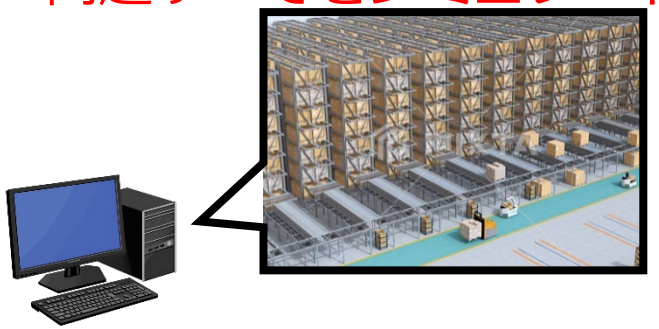
マルチエージェントシステムの種類

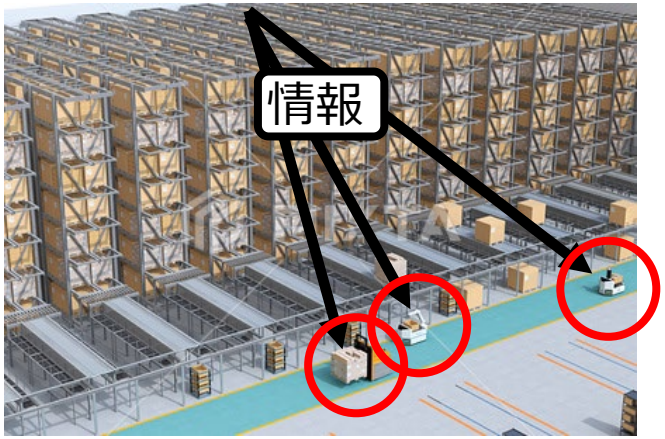
- 他エージェントの学習結果は得られない

本研究の対象

集中型マルチエージェント

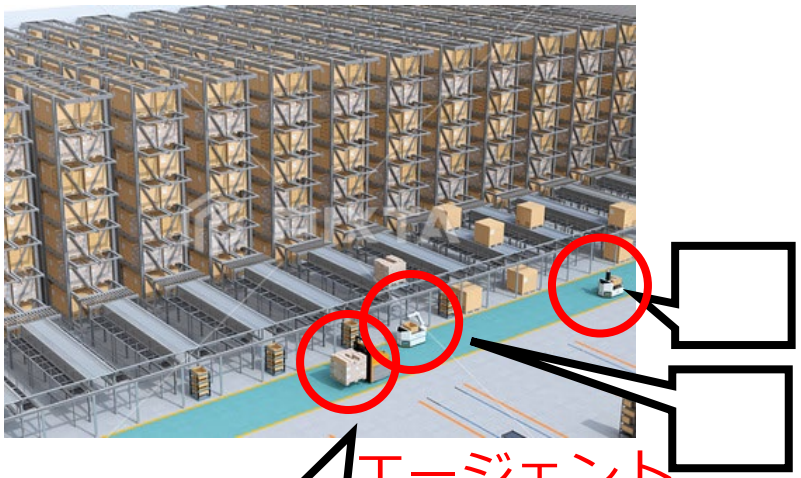
問題すべてをシミュレート





エージェント

分散型マルチエージェント



エージェント

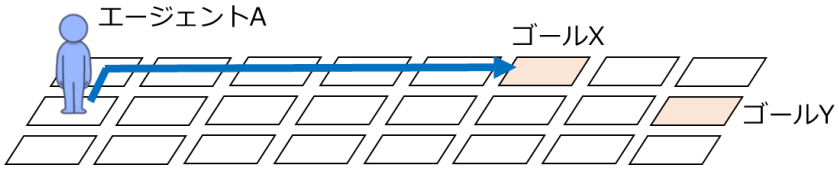
荷物を運ぶ

個々が目的を持つ

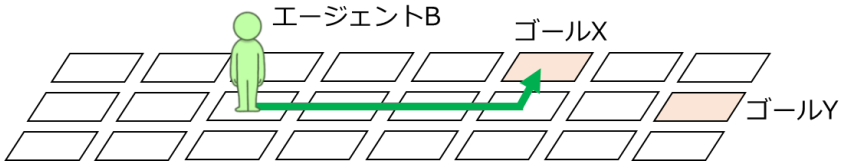
エージェントの協調

- 全体の目的≠個人の目的

⇒ 全体の目的達成のためにエージェントは協調



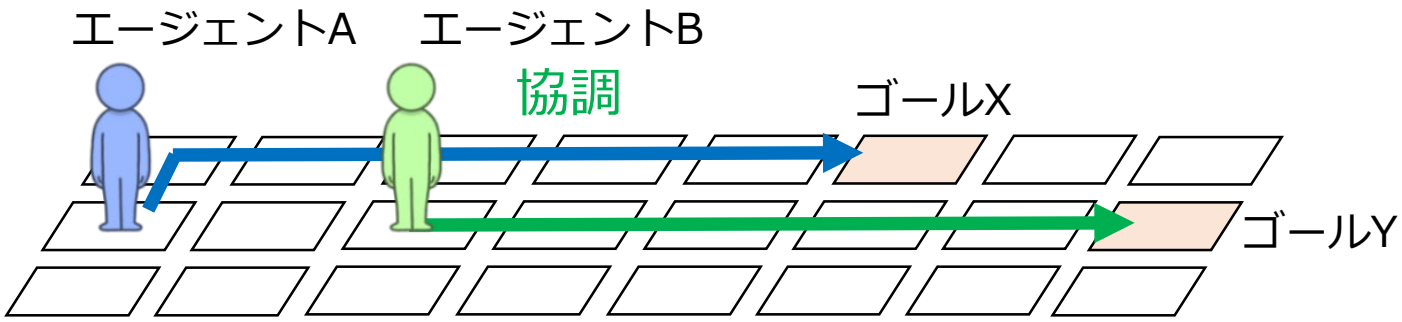
エージェントA(個人)の目的
→ 近くのゴールに到達



エージェントB(個人)の目的
→ 近くのゴールに到達

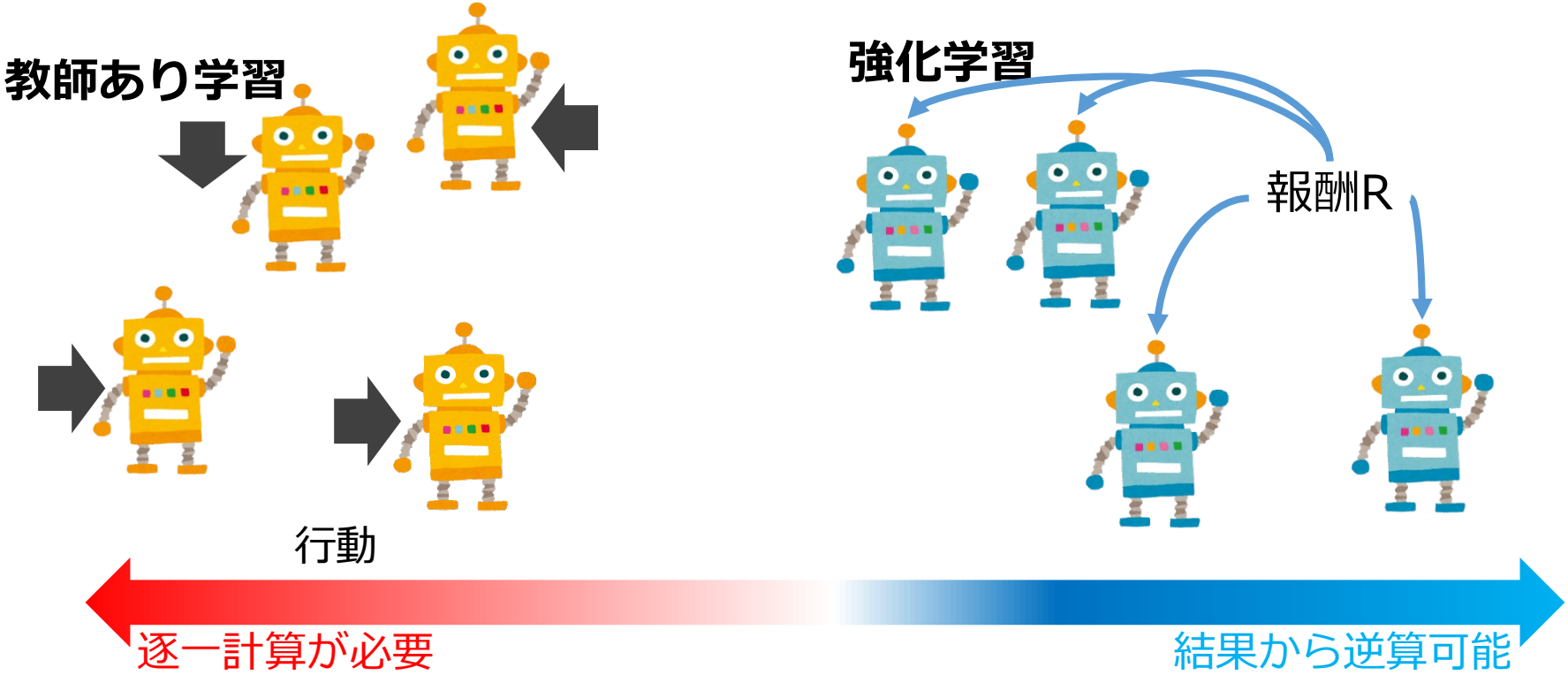
＝　≠

エージェント全体の目的



マルチエージェント強化学習
(Multi-Agent Reinforcement Learning: MARL)

- エージェント間の協調行動を学習
- 個人/全体の利益最大化のためのエージェント制御



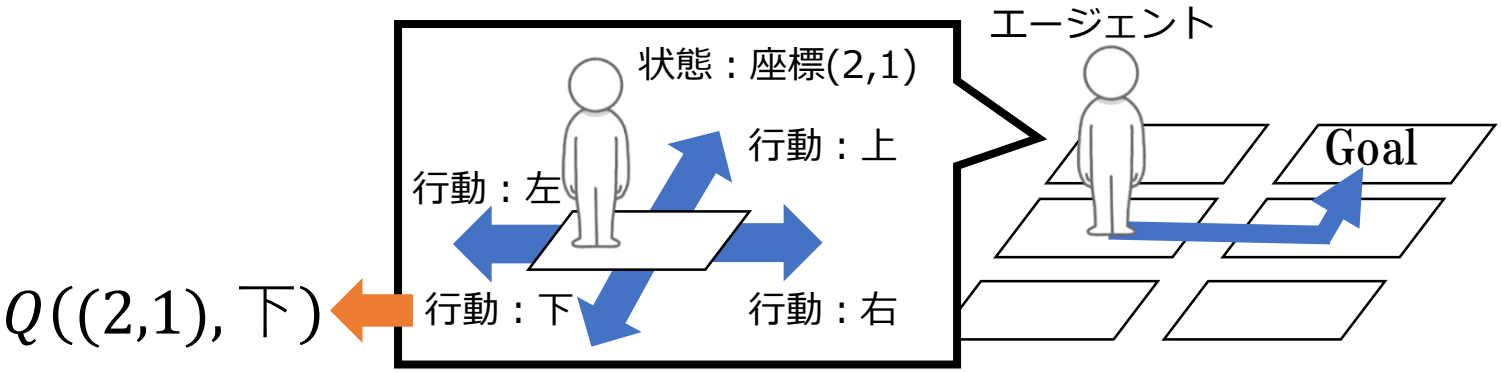
1つ1つの正解行動を与える必要あり

正確ではないが結果から学習可能

強化学習（Q学習の例）

- 経験に基づく機械学習手法
  - 目的さえ与えれば目的達成までの行動が学習可能
- 報酬に基づく入力に対する出力の価値推定
  - 入力：状態（位置）と報酬
  - 出力：行動（上下左右）
  - 報酬：学習の手掛かり

状態行動価値（Q値）：自身の状態と行う行動の期待報酬

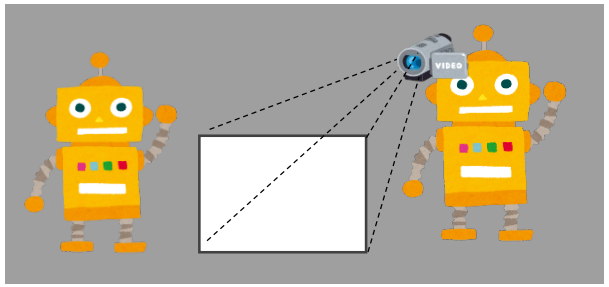


更新式： $Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)]$

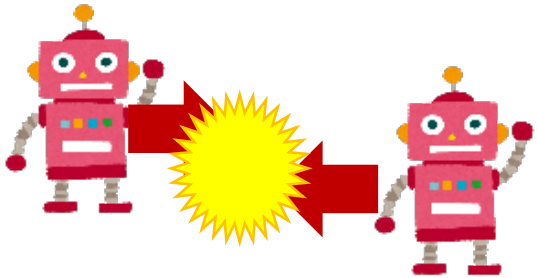
s: 現在の状態, a: 現在の行動, s': 次状態, a': 次状態行動, r: 報酬, α: 学習率[0,1], γ: 割引率[0,1]

従来手法の現状と問題点

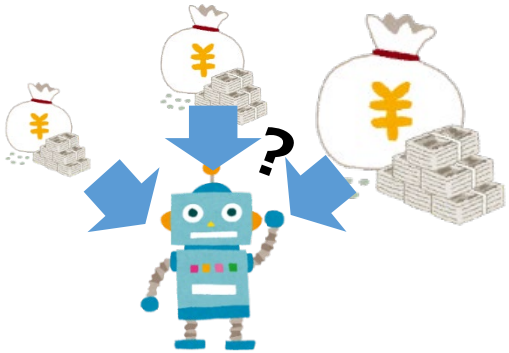
MASにおける強化学習の問題[Arai+, 2001]



不完全知覚問題



同時学習問題



報酬分配問題

認知

学習

目的

エージェント

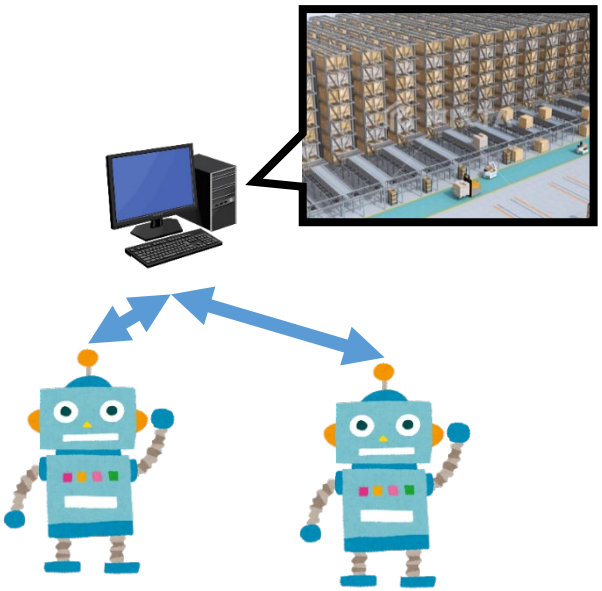
環境

MASのモデル化

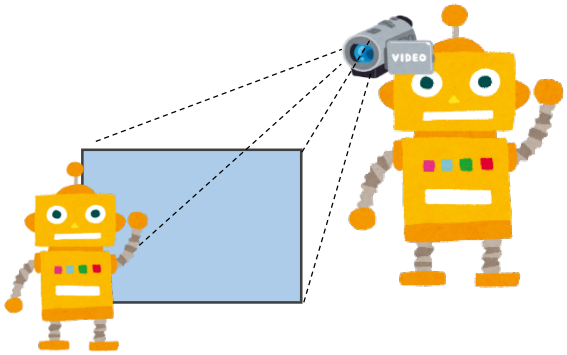
学習エージェントのモデル化

- 通信情報から報酬により行動ルールを学習
  - 全情報 (グローバル通信)：神様視点の情報
  - 近傍情報のみ (ローカル通信)：エージェント周辺の情報のみ
  - 情報なし (通信なし)：情報を利用しない

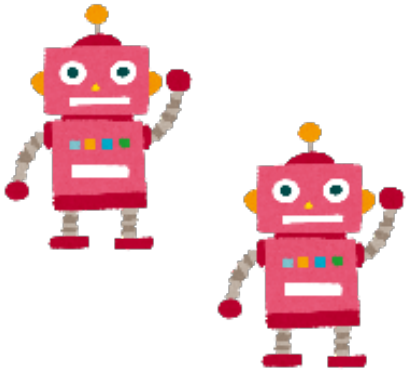
全情報



近傍情報のみ

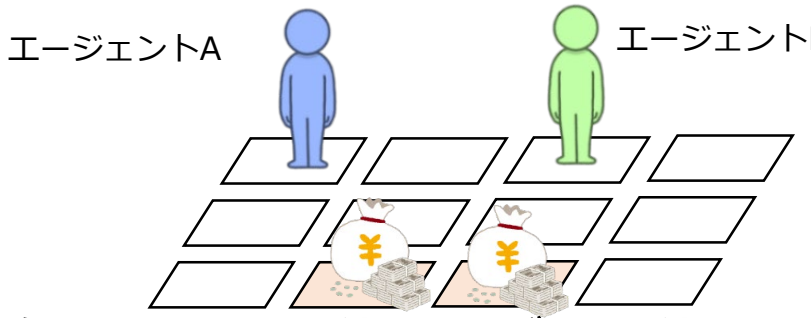


情報なし

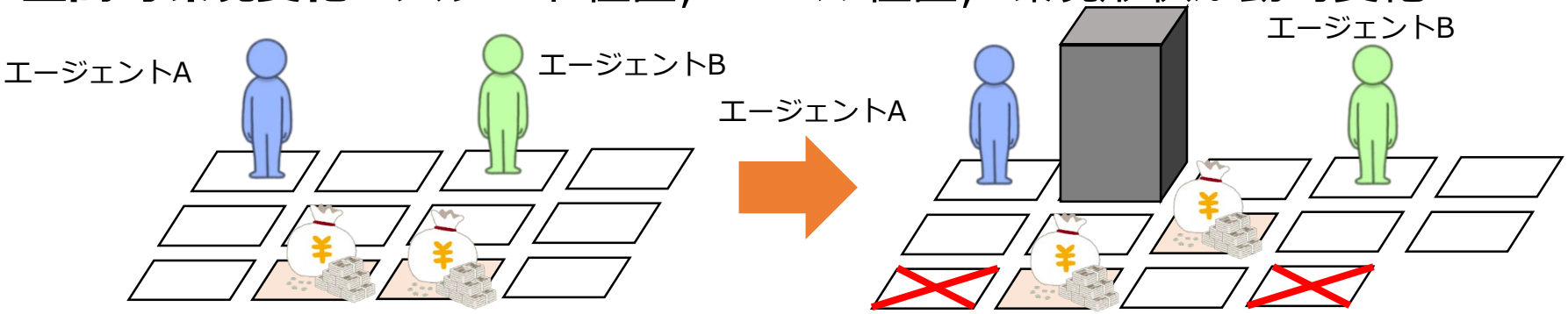


環境のモデル化（例：迷路問題）

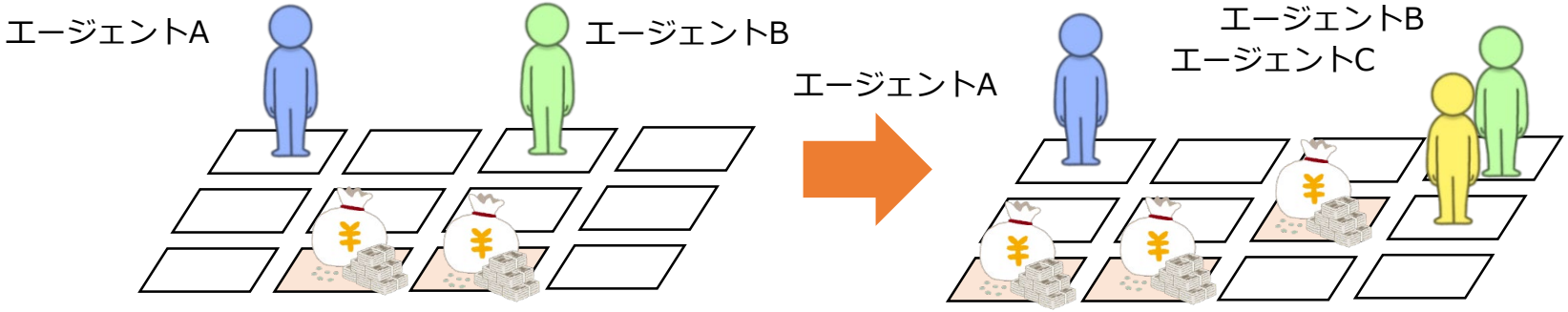
静的環境：エージェントの動作以外変化なし



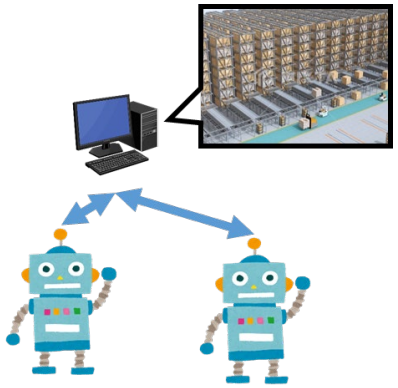
空間的環境変化：スタート位置，ゴール位置，環境形状が動的変化

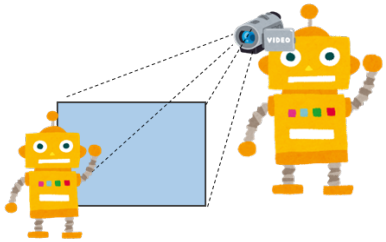


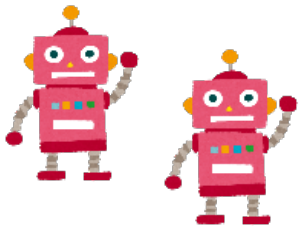
エージェント・ゴール数変化：エージェント数とゴール数が変化



従来研究と比較した位置付けの違い







易 → 難

	全情報 (グローバル通信)	近傍情報のみ (ローカル通信)	情報なし (通信なし)
静的環境	[Raileanu+, 2018] [Littman, 1994] [Devlin+, 2014]	[Shiraishi+, 2018]	
動的環境	[Egorov, 2016] [Verma+, 2019]	[Zemzem+, 2015] [Arai+, 2000]	

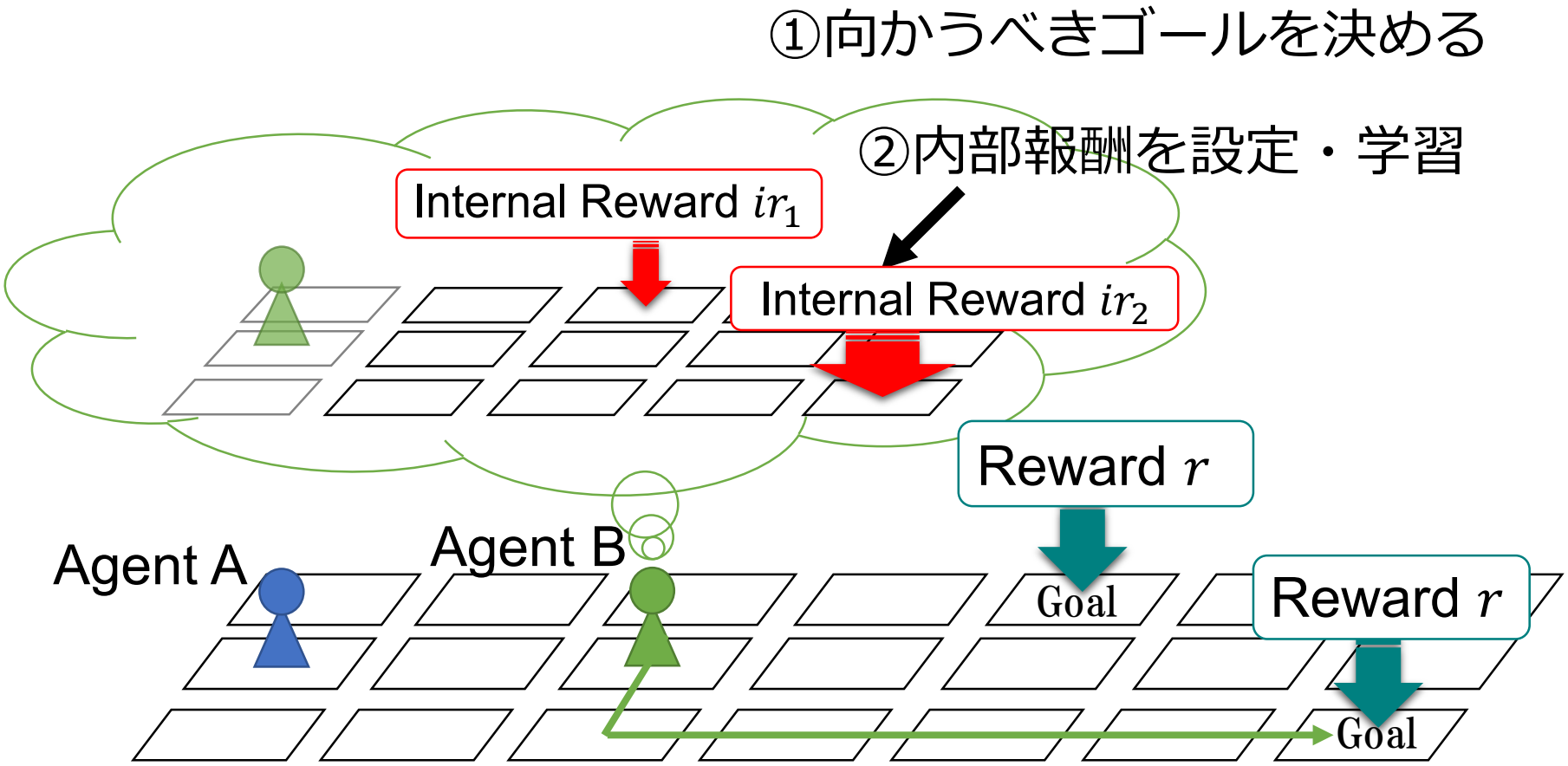
易 ↓ 難

従来法の想定する範囲

学生時代までの研究で想定した範囲

研究に対する自身の着眼点

環境の情報（特に報酬）のみを利用した最適制御

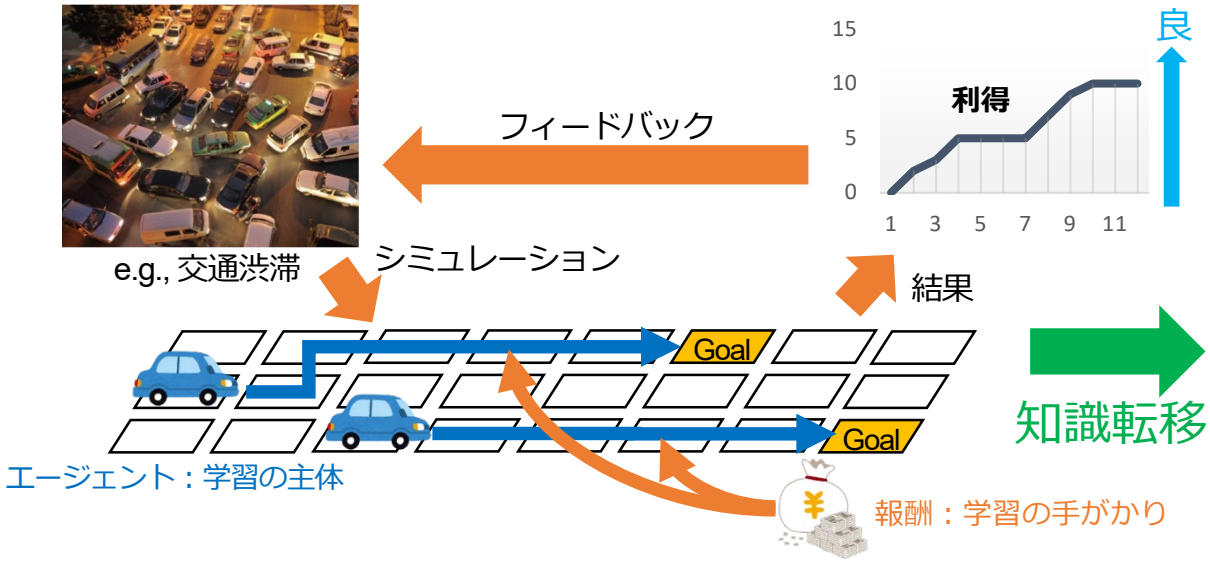


最近の話題

研究背景

- マルチエージェント強化学習

  → 強化学習による集団行動の獲得



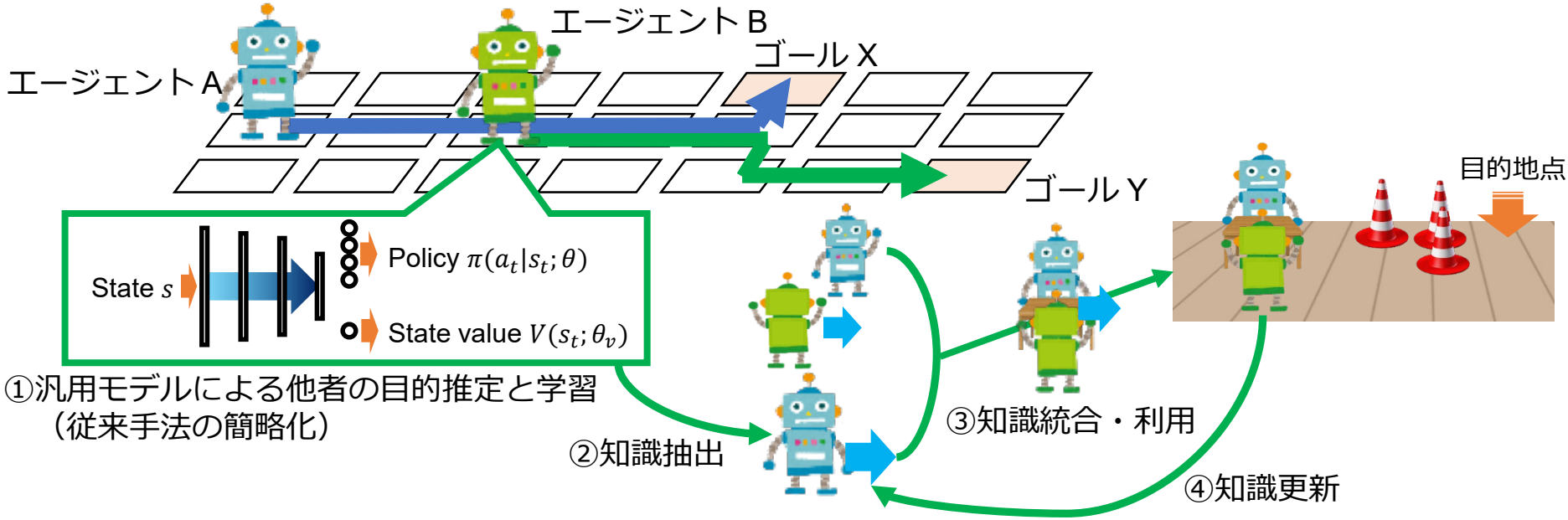
e.g., ロボット制御

- 他の環境への展開：学習した知識の生成と利用
- 知識の利用に関する従来研究
  1. 自身の学習器を利用した他者の目的推定 [Raileanu+, 2018]
  2. メタ学習による知識転移 [Frans+, 2018]

  → 未知環境への適用に課題

研究目的

- 未知環境に向けた知識生成とその利用



- 解決すべき課題

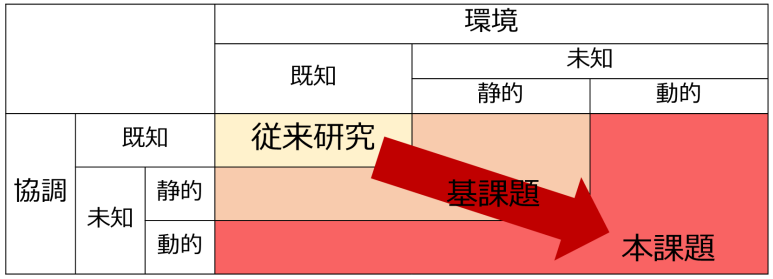
目的

① 汎用モデルによる他者の目的推定と協調行動学習

② 学習済みモデルからの知識抽出

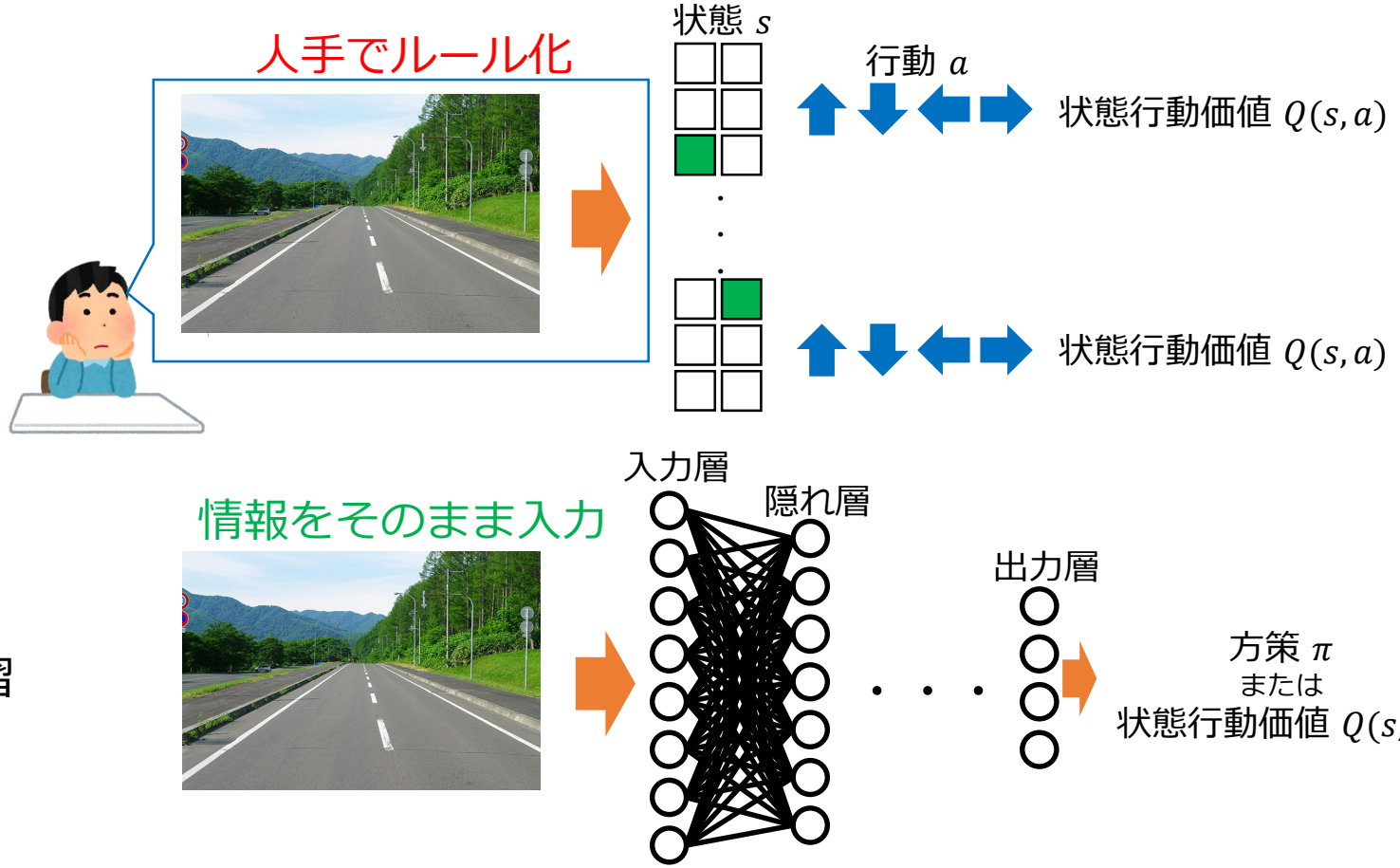
③ 抽出した知識の統合

④ 知識の更新



深層強化学習

- ニューラルネットワークで行動価値や方策を推定
  - 入力情報の抽象化と出力関数の近似
- 複雑な入出力関係でも学習可能



Asynchronous Advantage Actor-Critic(A3C) [Mnih+, 2016]

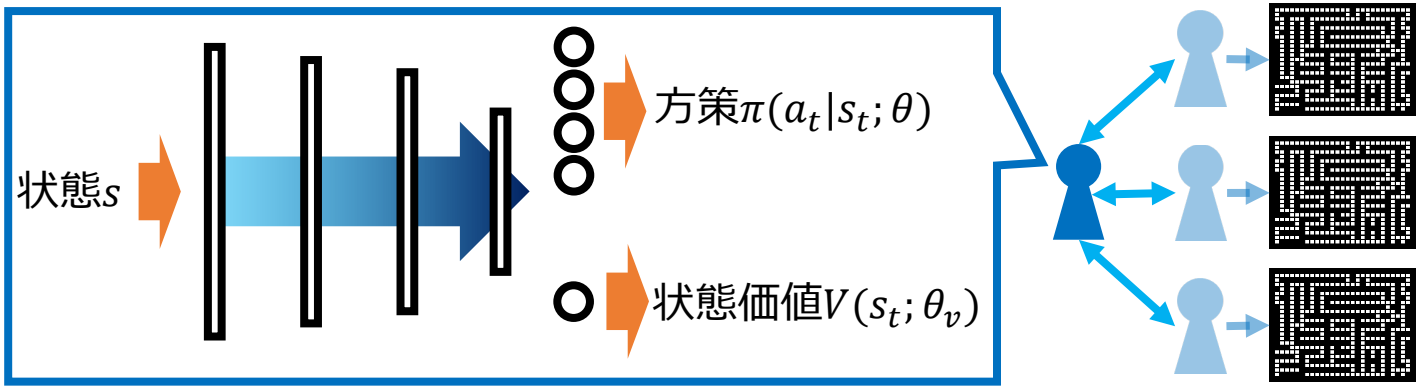
- 複製したエージェントの結果の統合：Asynchronous
- 数手先の状態を含めた学習：Advantage
- 方策と状態価値による安定した学習：Actor-Critic

$$d\theta \leftarrow d\theta + \nabla_{\theta'} \log \pi(a_i|s_i;\theta') A(s_i, a_i; \theta, \theta') + \beta \nabla_{\theta'} H\big(\pi(s_i;\theta')\big),$$

$$A(s_i, a_i; \theta, \theta_v) = \sum_{j=0}^{k-1} \gamma^j r_{i+j} + \gamma^k V(s_{i+k}; \theta_v) - V(s_i; \theta_v)\,.$$

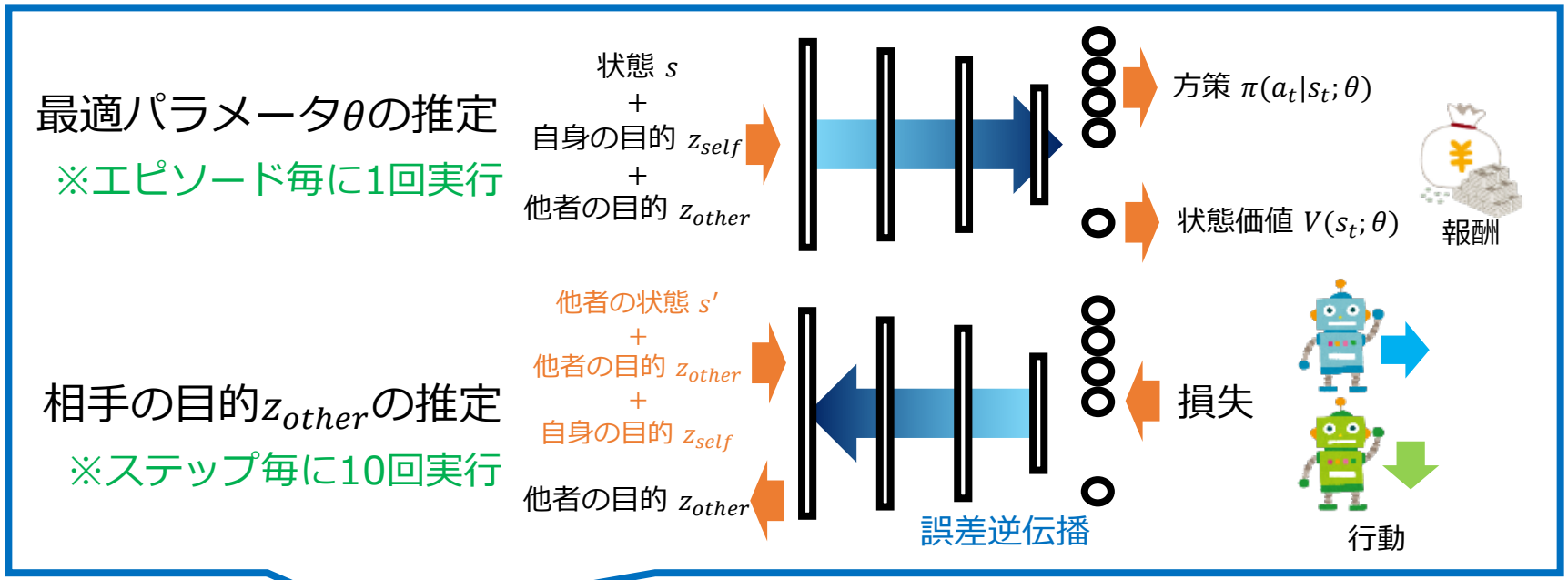
$$d\theta_v \leftarrow d\theta_v + \frac{\partial \big(R - V(s_i; \theta_v')\big)^2}{\partial \theta_v'}, \qquad R \leftarrow r_i + \gamma R.$$

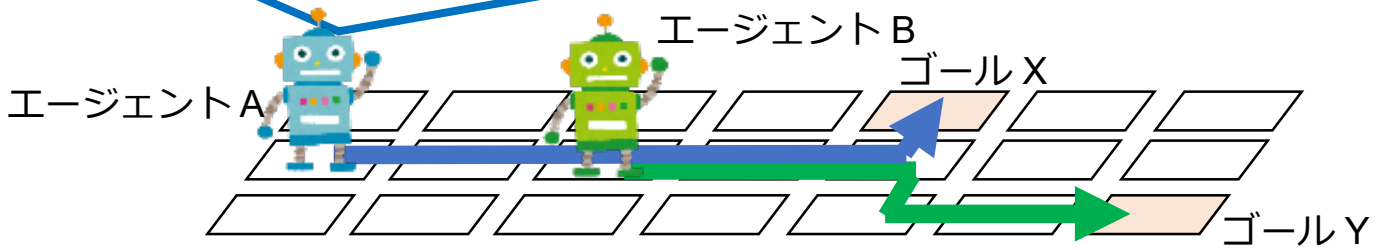
θ：ネットワークパラメータ,
θ'：複製エージェントのネットワークパラメータ,
i: エピソード内の任意のステップ数,
r_i:エピソードiの獲得報酬値



Self-other modeling (SOM) [Raileanu+, 2018]

- 推定した他エージェントの目的に基づく協調行動学習
  - 報酬から自身の学習器（A3C[1]）のパラメータを推定
  - 自身の学習器からの他エージェントの目的推定

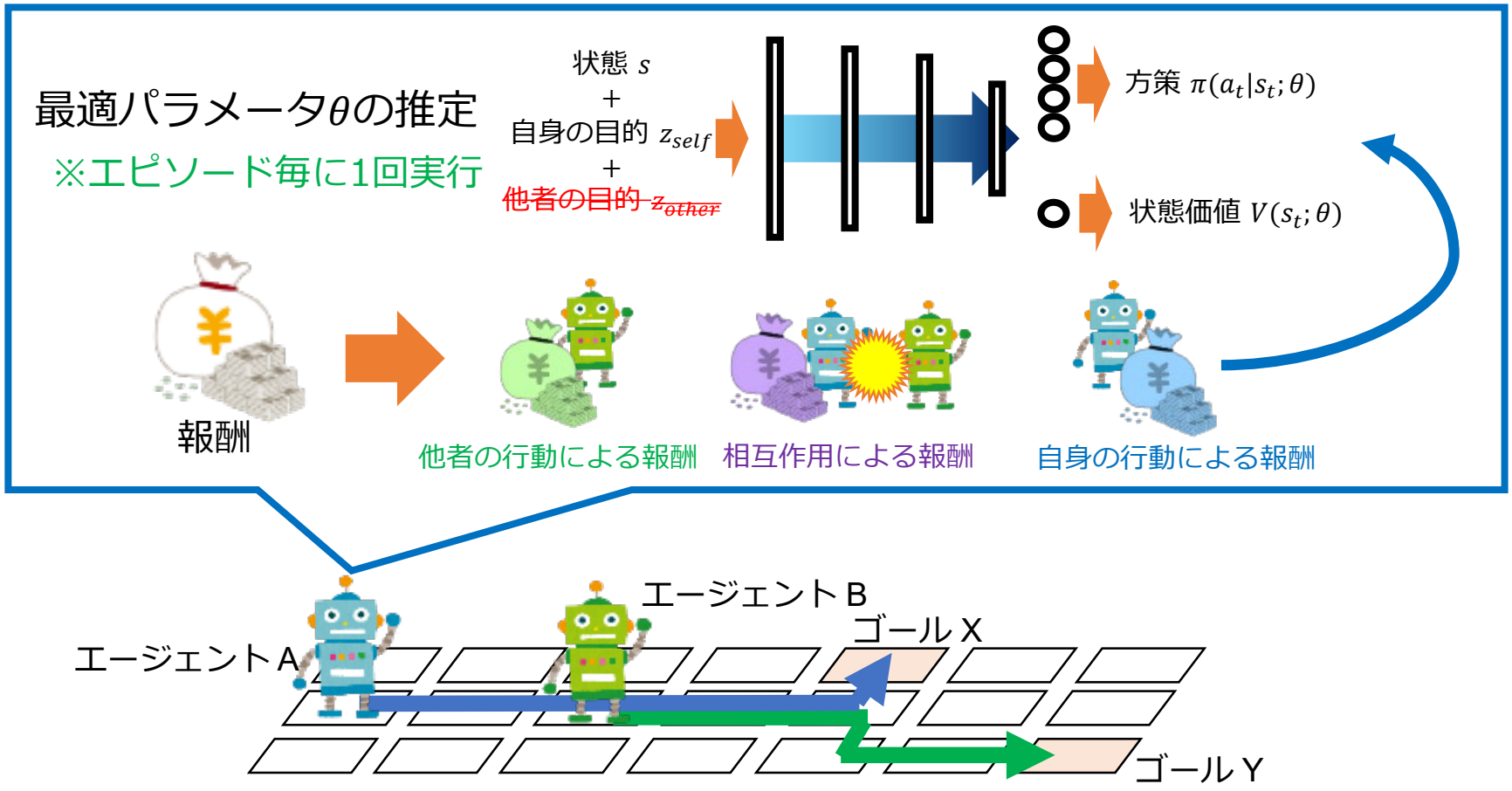




[1] Mnih, V., et al. “Asynchronous methods for deep reinforcement learning.” arXiv, 1602.01783, 2016.

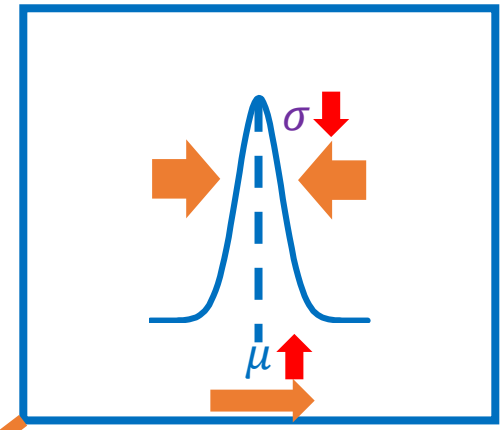
アプローチ

- 獲得報酬から他エージェントの影響を推定
- 他エージェントに影響を与えない報酬値から学習



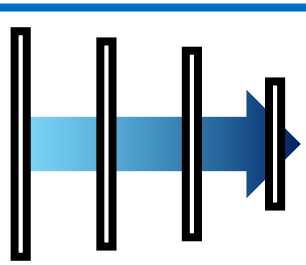
提案手法

- 各ゴール正規分布に従う報酬関数を仮定
  - $f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
- 獲得報酬からパラメータ（μ, σ）を推定
- μの期待値を内部報酬値に設定
  - $r = \frac{\mu}{\sqrt{2\pi}\sigma}$



理想の報酬関数

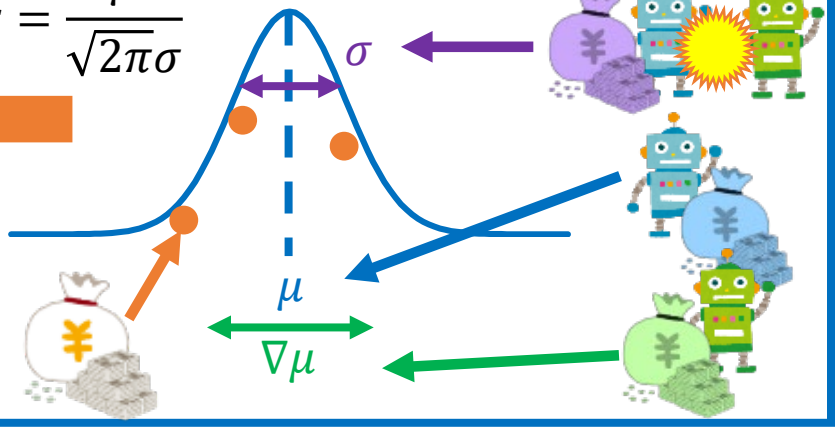
状態 s
＋
自身の目的 z_{self}
＋
~~他者の目的 z_{other}~~



方策 $\pi(a_t|s_t;\theta)$

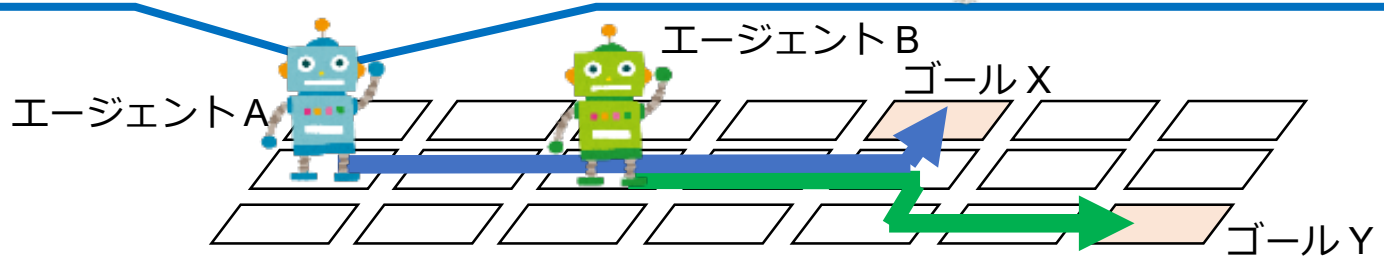
状態価値 $V(s_t;\theta)$

$r = \frac{\mu}{\sqrt{2\pi}\sigma}$



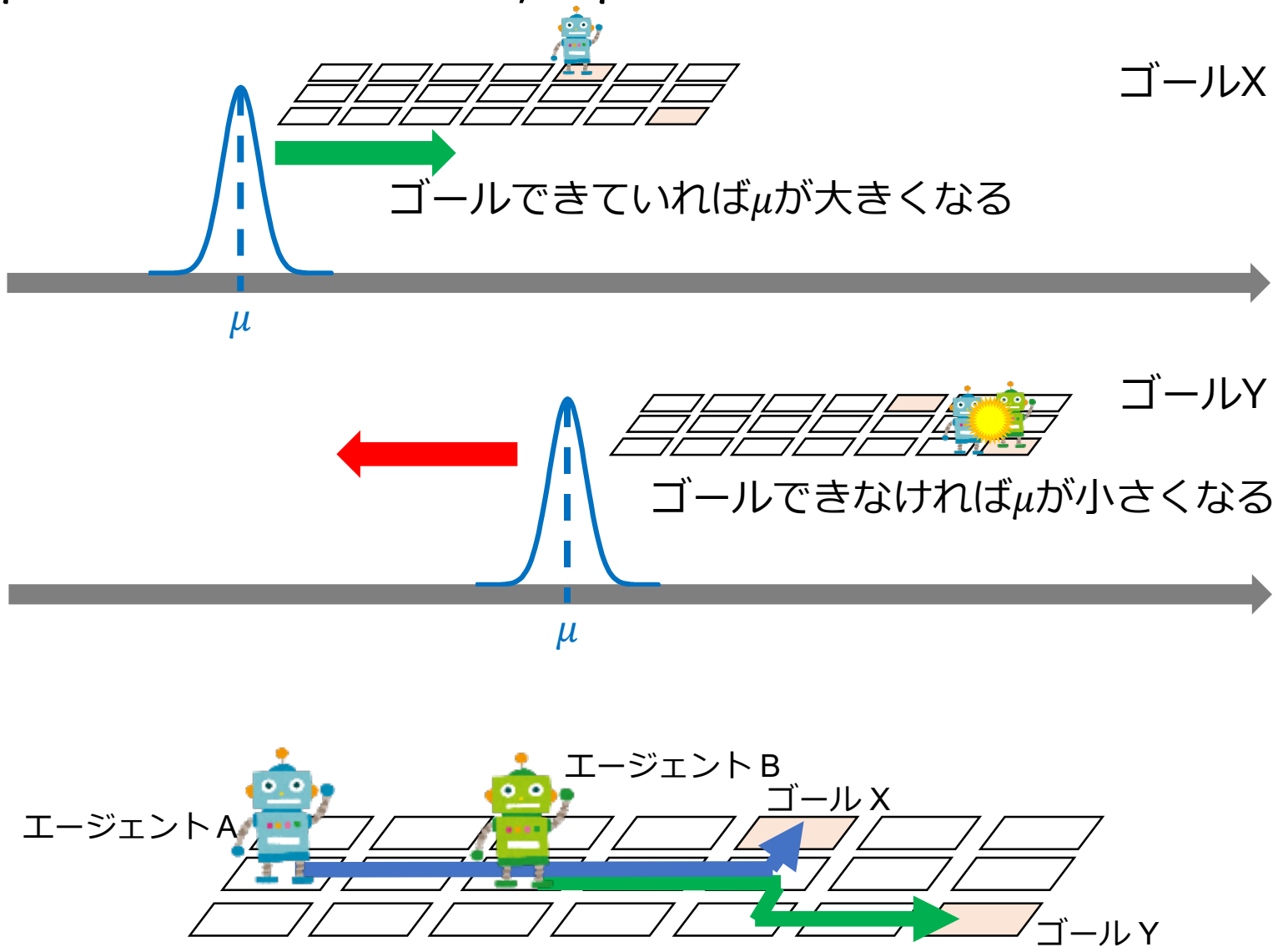
最適パラメータθの推定

※エピソード毎に1回実行



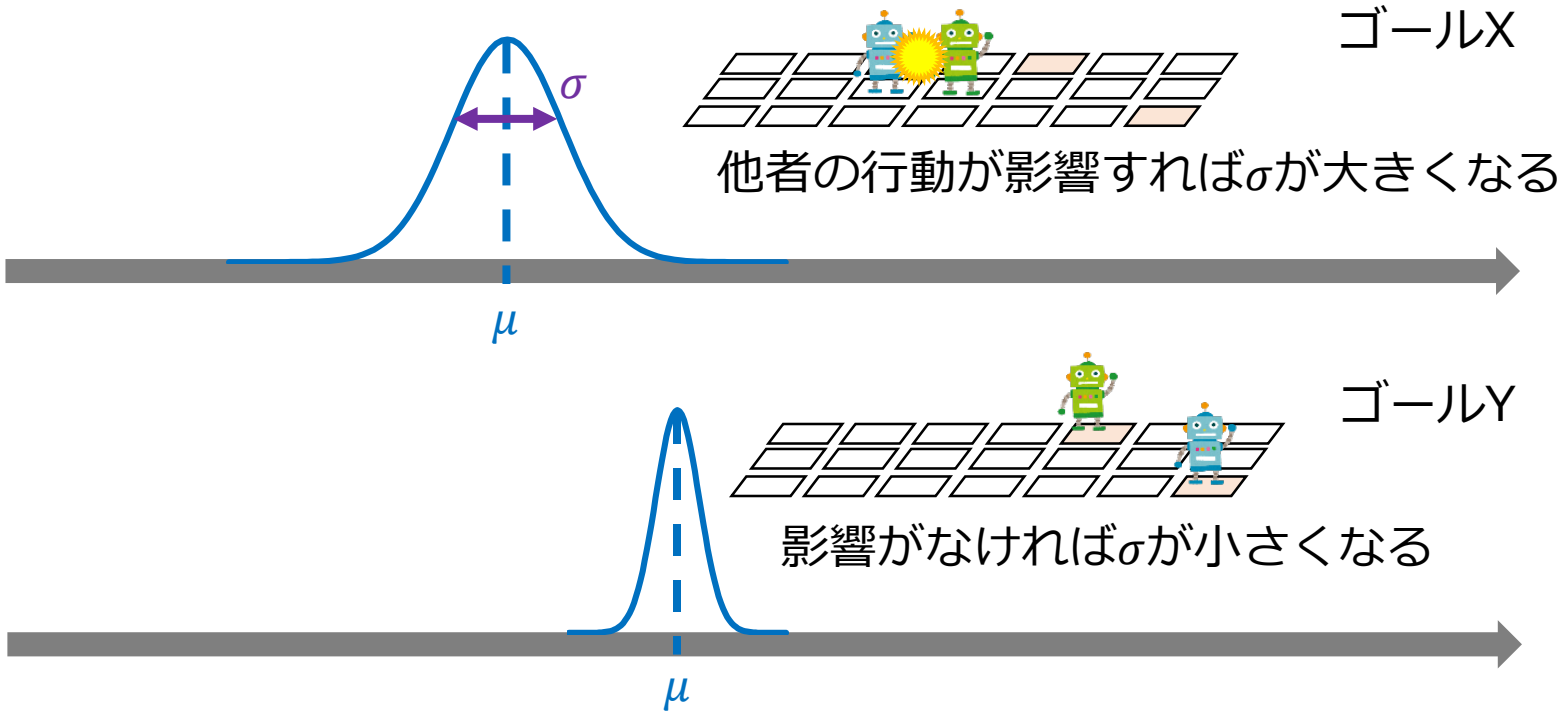
報酬関数のイメージ(μ, $\nabla\mu$)

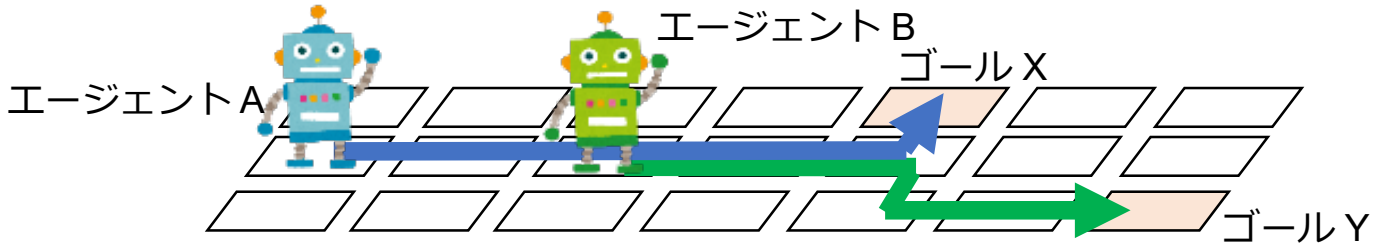
- μは自身の行動の結果，$\nabla\mu$他者の行動の結果



報酬関数のイメージ(σ)

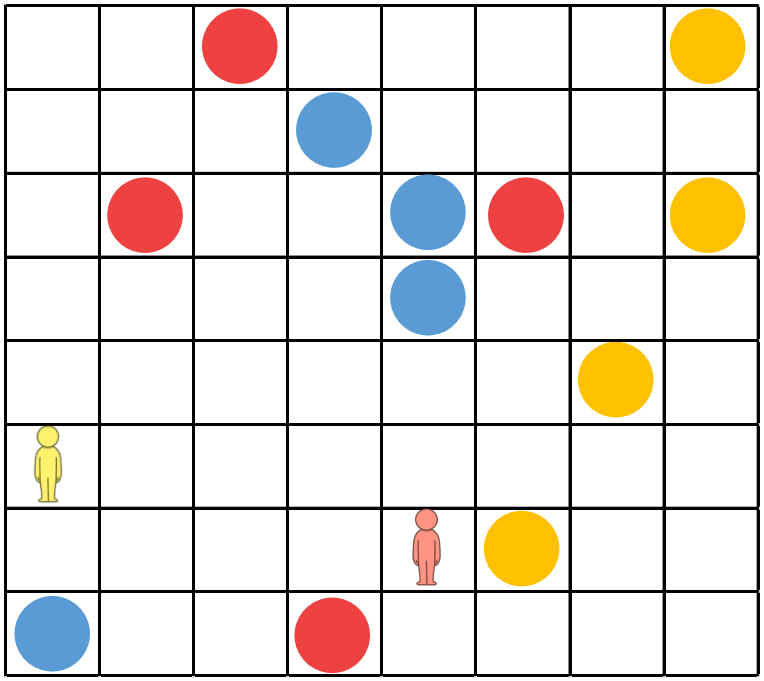
- μは自身の行動の結果，∇μ他者の行動の結果





実験内容

- 提案手法の性能をSOMおよびA3Cと比較
- 問題：Coin Game [Raileanu+, 2018]
- 評価：獲得報酬値
  - 提案手法は10000エピソードの移動平均の結果を表示
- パラメータ
  - エピソード数：20,000,000
  - 最大ステップ数：10
  - プロセス数：16
  - 学習率α：0.0007
  - 割引率γ：0.99
  - 定数β：0.01



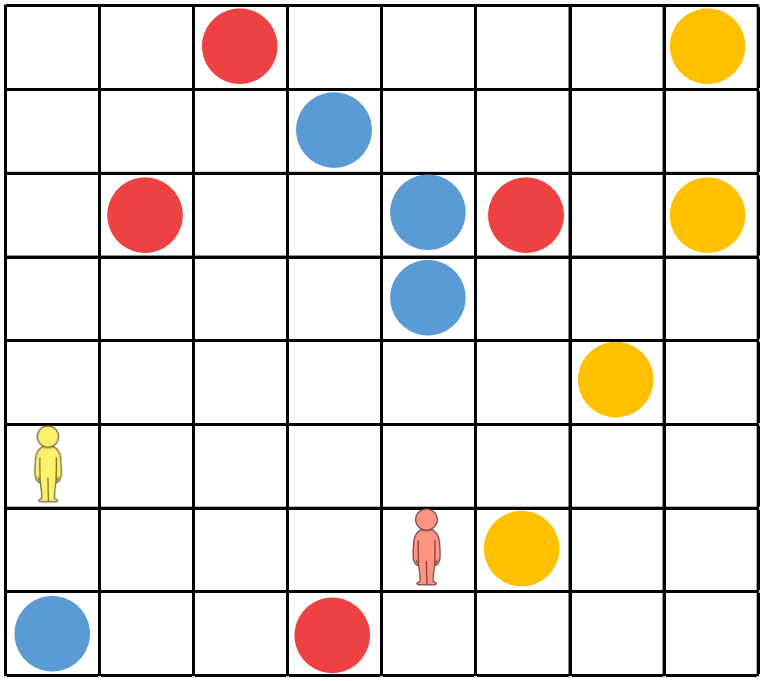
Coin Gameの概略図

Coin Game [Raileanu+, 2018]

- 決められた色のコインの獲得を目指す問題
  - 相手の色のコインを取らず自身の色のコインの獲得を目指す
- エピソード毎に環境がランダムに変化
  - エージェントのスタート位置
  - コインの位置
  - 各エージェントが獲得すべきコインの色

$$R_{coin} = \left(n_{C_{self}}^{self} + n_{C_{self}}^{other}\right)^2 + \left(n_{C_{other}}^{self} + n_{C_{other}}^{other}\right)^2 - \left(n_{C_{neither}}^{self} + n_{C_{neither}}^{other}\right)^2$$

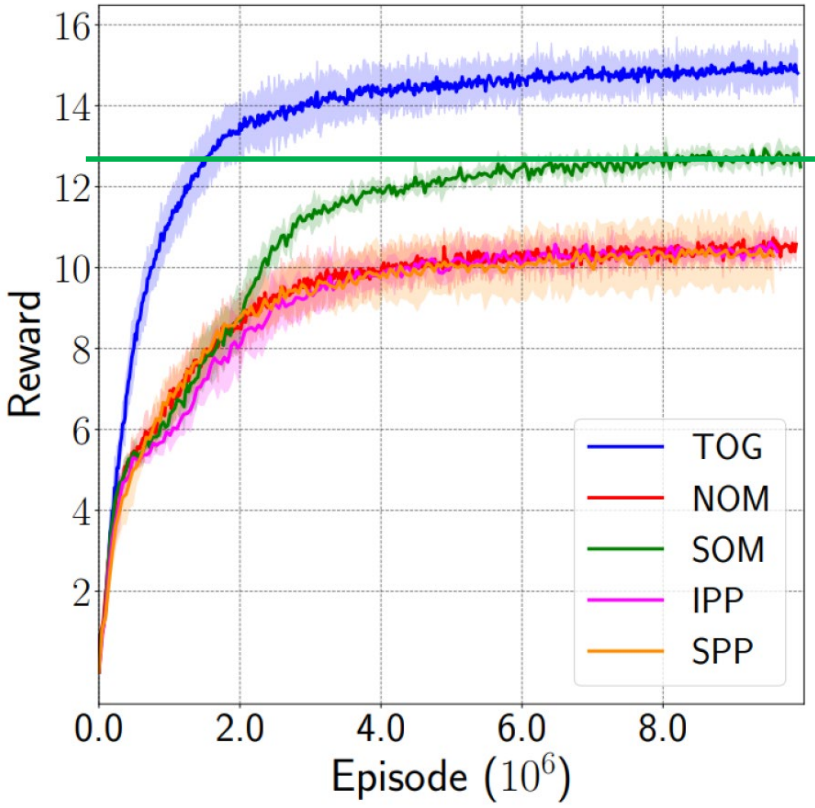
R_{coin} ：両エージェントの獲得報酬値
$n_{C_{self}}^{self}$ ：自身の色のコインを獲得した数
C_{self} ：自身の色のコイン
C_{other} ：他者の色のコイン
$C_{neither}$ ：両者にない色のコイン

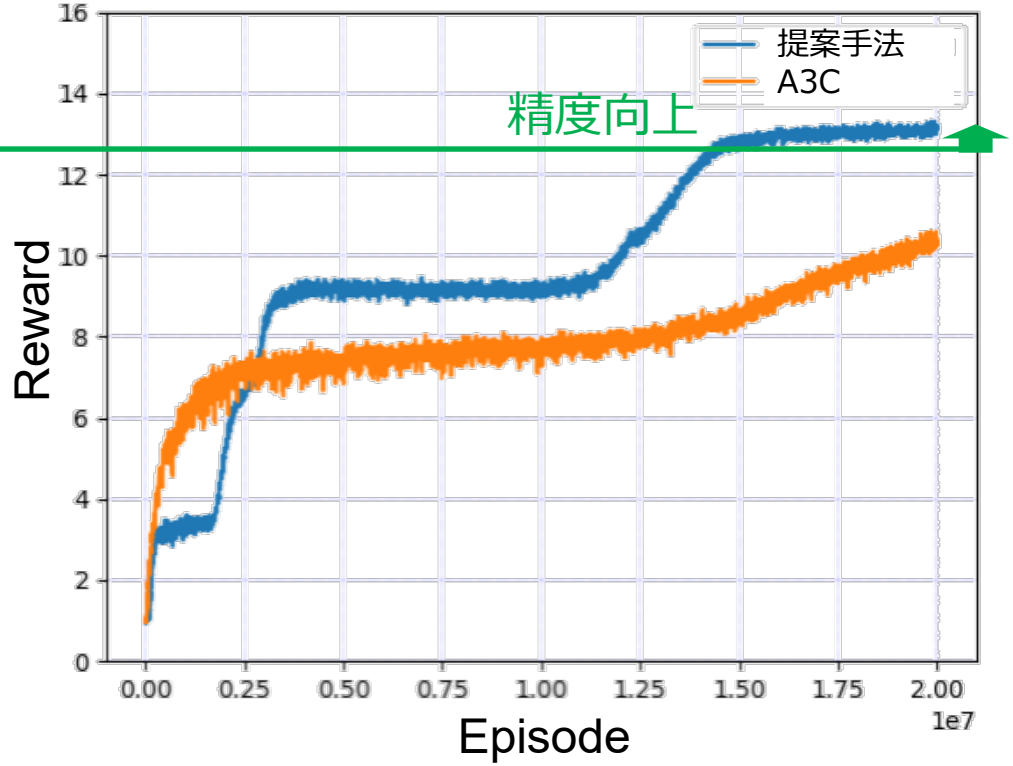


Coin Gameの概略図

実験結果（SOMおよびA3Cとの比較）

- 提案手法のほうが精度の面で上回る
- エピソードに10倍の差があるが学習回数の規模は同等
  - SOMは毎ステップで10回相手のゴールを推定するため
  - 提案手法のほうがより早く最高精度に収束



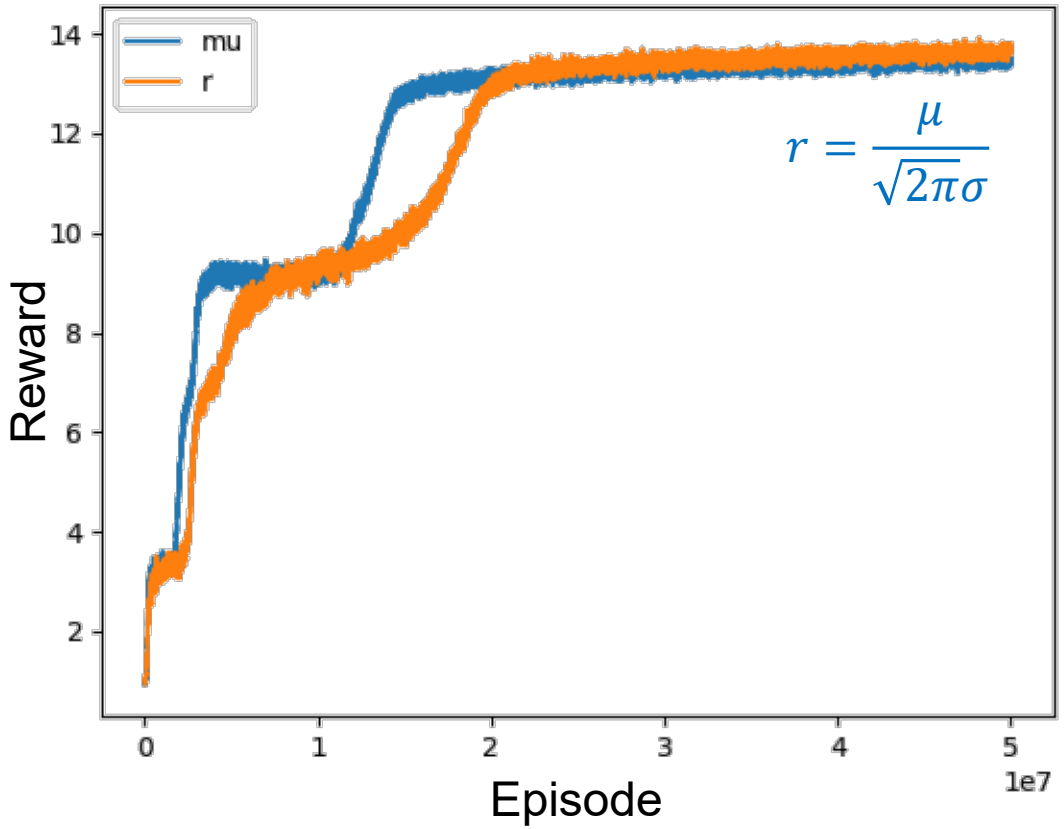


SOMの精度（文献[2]より引用）

[2] Raileanu, R., et al. “Modeling others using oneself in multi-agent reinforcement learning.” In Proc. of ICML 2018, pp. 4257-4266, July, 2018.

実験結果（報酬関数同士の比較）

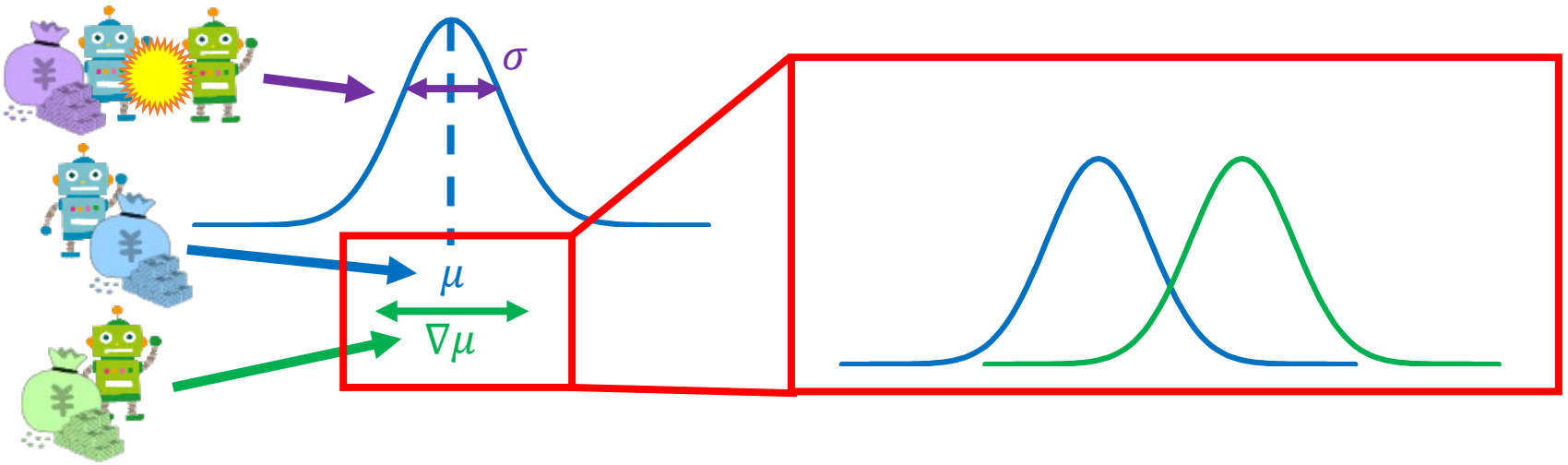
- 平均値μと報酬値rの性能の違い
  - 平均値では収束が早い
  - 報酬値では収束は遅いが平均値よりも僅かに精度が上回る
    - 学習の収束後は報酬値の方が適切に各エージェントの方策を評価できているため



報酬関数近似に関する考察

- 正規分布の仮定はある程度正しく報酬関数を近似可能
- 提案手法では最適行動を学習しなかった
  - 獲得したコインの数はSOMと大きく変わらない
  - 正規分布の仮定では他エージェントの獲得報酬の要素が不足

自身の獲得報酬と他者の獲得報酬に関してより詳細な仮定が必要

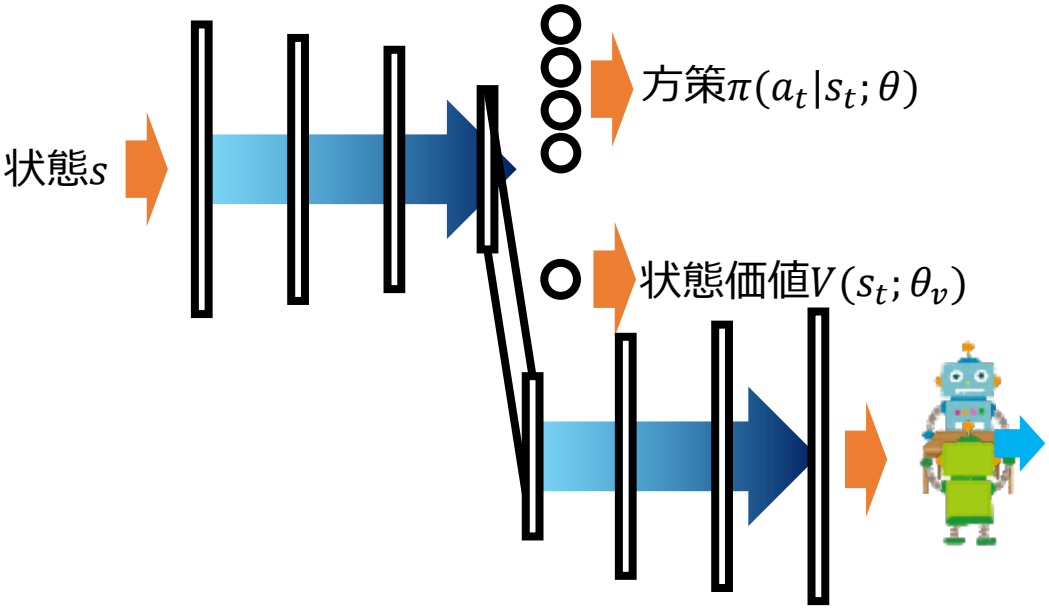


まとめ

- 研究目的
  - 報酬関数近似のみを用いた暗黙的協調行動学習法の提案
- 提案手法
  - 獲得報酬関数モデルを仮定
  - 相互作用のノイズを最小化するように学習
- 実験
  - Coin GameにおいてSOMよりも高精度に学習
  - SOM同様最適方策は獲得せず
- 今後の課題
  - 仮定の検証：正規分布の妥当性
  - より適切な分布の探求：他のエージェントの要因の近似推定
  - 他のテスト問題での検証：レシピゲーム，ドアゲーム，他

今後の展望

1. 学習結果からの知識の抽出と利用



2. 未知環境の動態に適応した知識の進化