グループ 24:複数の世界モデルを用いた意思決定

大塚　崇明[*1]　　菊永　陸斗[*2]
Takaaki Ohtsuka　　Rikuto Kikunaga

[*1] 放送大学　　[*2] 聖マリアンナ医科大学
The Open University of Japan　　St. Marianna University School of Medicine

We explore how multiple agents with different values can cooperate to make optimal decisions. We focus on multi-agent reinforcement learning, with particular attention to algorithms such as MADDPG and MATD3. These algorithms are suitable for learning in a multi-agent environment and available in AgileRL. In this study, the MATD3 algorithm was applied in 9 different environments using PettingZoo MPE (Multi Particle Environments). We trained 100,000 episodes in each environment and analyzed the results along with the trends in episodic rewards. In particular, we observed that the stability of the learning curve differed when the agent had a single objective and when it was influenced by multiple conditions. The results suggest that the more multiple factors influence the agent's behavior, the more unstable the learning becomes. Future research will involve creating a custom multi-agent environment in PettingZooMPE to separately assess learning behaviors of benign and hostile agents.

1. 研究背景・目的

我々は，ある問題を取り組む際に，複数人で相談しながら最終的な意思決定を行うことが多い．これは，それぞれが異なる価値観（世界モデル）を持っており，それらを組み合わせることで，

最終的により良い意思決定をすることを目指しているからである．本プロジェクトでは，複数の環境やタスクで学習した世界モデルがあると仮定し，それらの予測や推論を適切に利用することで新しい問題に対処できる方法を検討する．

2. 関連研究

マルチエージェント強化学習を実現するための強化学習アルゴリズムと、複数のエージェントの活動を仮想的に実現した環境がそれぞれ公開されていた。

2.1 強化学習アルゴリズム

マルチエージェント環境での強化学習を行うためのアルゴリズムとして、過去に下記の手法が提案されていた。

- MADDPG　(2017) [1]
- MATD3　　(2019) [2]
- SA-MATD3 (2021) [3]

とくに、MADDPG および MATD3 は強化学習のための Python ライブラリ AgileRL [4] 内の機能のひとつとして組み込まれており、比較的簡単に利用ができる環境が整備されていた。

2.2 MPE 環境

アルゴリズムとは別に、マルチエージェントが相互作用を及ぼしながら活動するための仮想環境が必要となる。前述のアルゴリズムの実験・検証に利用された MPE 環境（MPE：Multi Particle Environments）は、継続したメンテナンスが実施されており、現在では PettingZoo [5] と呼ばれる Python ライブラリに管理が移管され、公開されている状況であった。

3. 提案手法（取り組んだこと）

PettingZoo MPE には、9 種類のマルチエージェント環境が提供されている。この、9 つの MPE 環境それぞれに対して、MATD3 アルゴリズムによるマルチエージェントの強化学習が実施できるように環境構築を行なった。そして、各環境において、10 万エピソードの学習を行い、結果を確認した。

4. 実験・考察

P2 から P6 にかけ、下記の項目で、MPE 環境の概要説明と学習結果を記す。ここで、エピソード報酬とは、そのエピソードでのエージェントの得た報酬の総和を指す。

- 環境の名称
- エージェント数
- 各点の意味
- 環境の概要説明
- 環境の描画例図
- 10 万エピソード学習時のエピソード報酬の推移図
- 学習結果についての考察

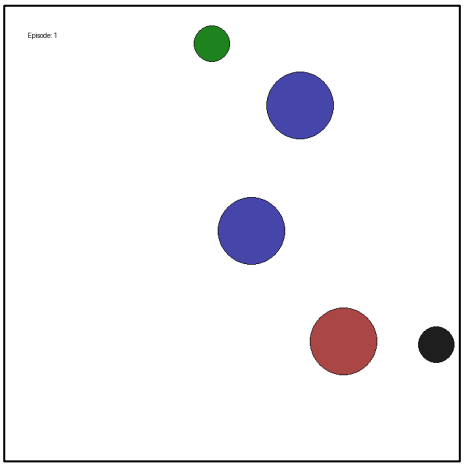
simple_push_v3, simple_reference_v3, simple_v3 等、環境内を移動するエージェントの目的が単一の環境においては学習曲線が比較的安定しやすいことがわかる。一方で simple_crypt_v3, simple_tag_v3, simple_world_comm_v3 等、敵性エージェントの行動がエージェントないしは良性エージェントの行動に影響を与える環境における学習曲線は不安定になりやすい傾向が確認できた。前者と後者の最大の違いは、エージェントが行動する際の目標が前者は単一であるのに比べ、後者は複数の条件から影響を受け、行動を更新する必要があることである。例えば、simple_push_v3 において良性エージェントが得る報酬を最大化するには、良性エージェントが可能な限りランドマークに近づけるように学習するだけで十分である。一方で、simple_push_v3 等のあるエージェントが他エージェントからの影響を受けるモデルでは、良性エージェントの報酬を最大化するには良性エージェントは障害物の位置および敵性エージェントの位置という複数の行動する環境上の要素から影響を受け、敵性エージェントから逃げ続けなければならない。以上のことから、影響を受ける要素の種類が増えるごとに学習の不安定性が増すことが予想される。より効果的な学習を実現するためには、複数の条件を包括する新たな評価軸を設定することが考えられる。

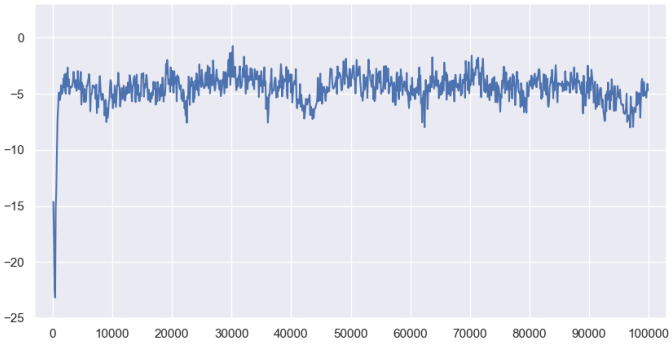
(1)　simple_adversary_v3

エージェント数：3

良性エージェント：青色
敵性エージェント：赤色
ターゲット・ランドマーク：緑色、灰色

赤色の敵性エージェントは、緑色か灰色かのどちらかに設定されたターゲット・ランドマークに近いほど正の報酬を得られる。青色の良性エージェントは、設定されたターゲット・ランドマークに近ければ正の報酬が得られるが、赤色の敵性エージェントがターゲット・ランドマークに近いほど負の報酬を与えられる。



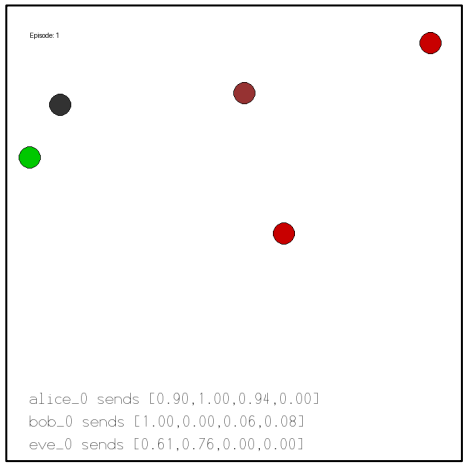


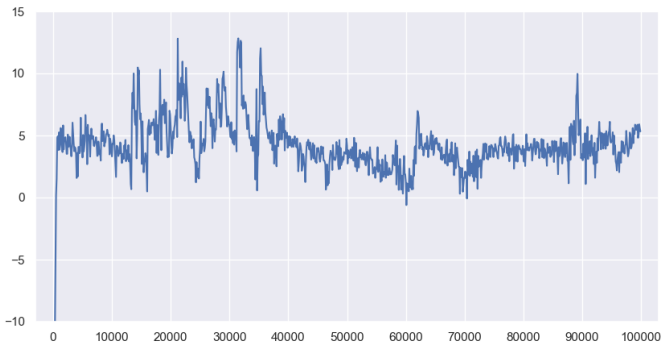
(2)　simple_crypto_v3

エージェント数：3

良性エージェント：アリス
良性エージェント：ボブ
敵性エージェント：イブ

この環境では、各エージェントは移動しない。
アリスは、ボブに 1bit のメッセージを送信する。ボブがアリスのメッセージを正しく受け取ることができたとき、アリスとボブは正の報酬が得られるが、イブに正しいメッセージを傍受されたとき、負の報酬を与えられる。アリスとボブはエピソードの初めにランダムに生成される秘密鍵を共有しており、この秘密鍵を利用してメッセージを送受信することを学ぶ。



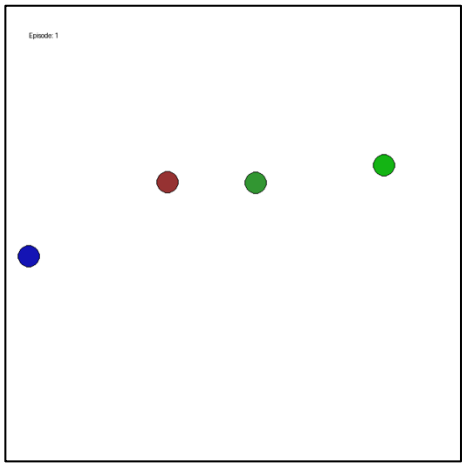


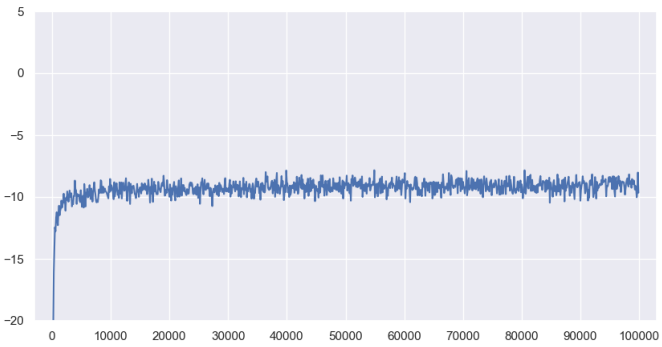
(3)　simple_push_v3

エージェント数：2

良性エージェント：目的のランドマーク色
敵性エージェント：赤色
ターゲット・ランドマーク：青色 or 黄緑色

良性エージェントは設定されたランドマークに近いほど正の報酬を得られる。敵性エージェントは、良性エージェントが設定されたランドマークから遠いほど正の報酬を得られる。敵性エージェントが良性エージェントを追い払うような動きが学習される。



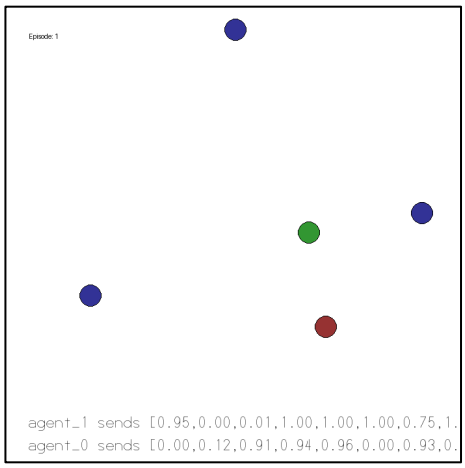


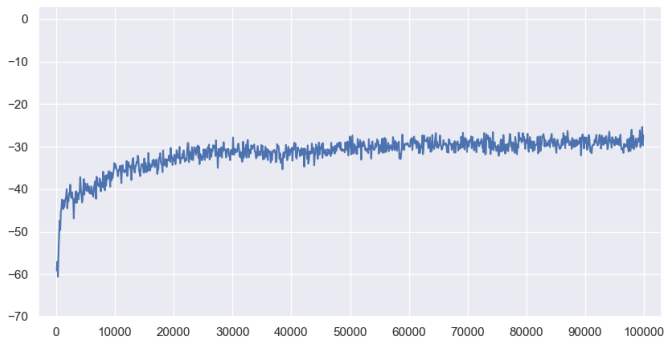
(4)　simple_reference_v3

エージェント数：2

エージェント1：目的のランドマーク色
エージェント2：目的のランドマーク色
ターゲット・ランドマーク：赤色、青色、緑色

3つのターゲット・ランドマークがあり、2つのエージェントはそれぞれ目的とするターゲット・ランドマークが設定されている。目指すべきターゲット・ランドマークは、もう一方のエージェントが知っており、自分でないエージェントから送信される情報を受け取り、どのように移動するか決定する。各エージェントの報酬は、おのおのに設定されたターゲット・ランドマークにどれだけ近いかによって与えられる。



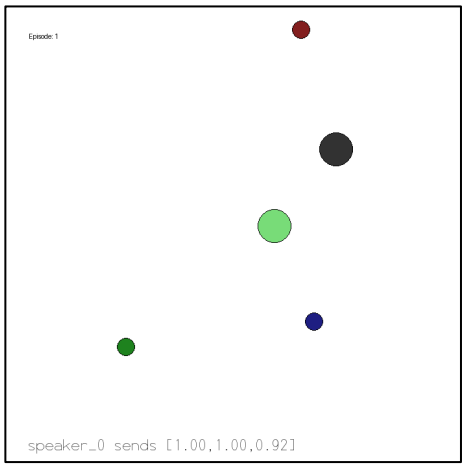


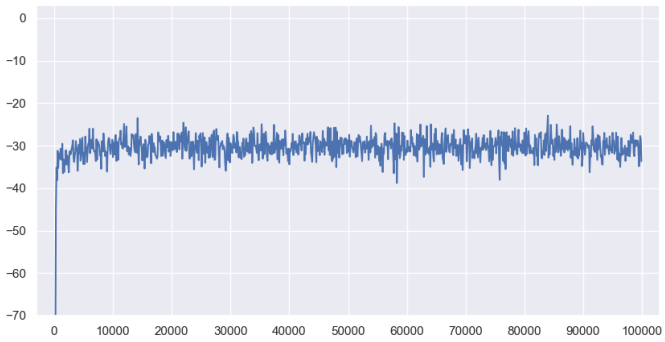
(5)　simple_speaker_listener_v4

エージェント数：2

スピーカー：灰色
リスナー：黄緑色
ターゲット・ランドマーク：赤色、青色、緑色

スピーカーは移動せず、リスナーに目指すべきターゲット・ランドマークの位置の情報を送信する。リスナーはスピーカーから送信された情報を受け取り、目的のターゲット・ランドマークへ移動する。



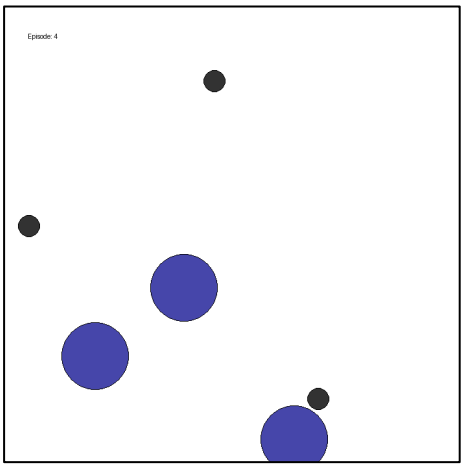


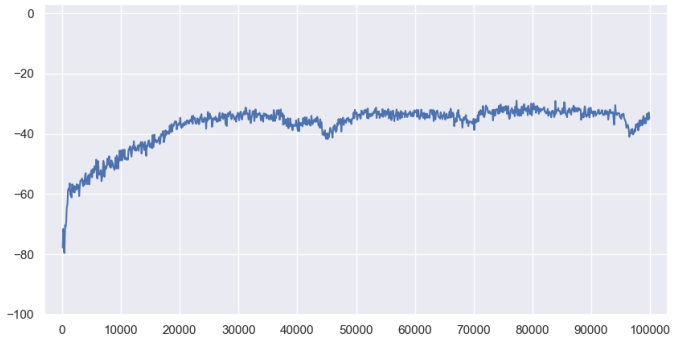
(6)　simple_spread_v3

エージェント数：3

エージェント：紺色
ターゲット・ランドマーク：灰色

3つのエージェントは3つのターゲット・ランドマークに近づくほど報酬を得られる。ただし、各エージェントが衝突すると負の報酬が与えられる。このため、各エージェントがそれぞれ別のターゲット・ランドマークを目指して移動するよう学習される。



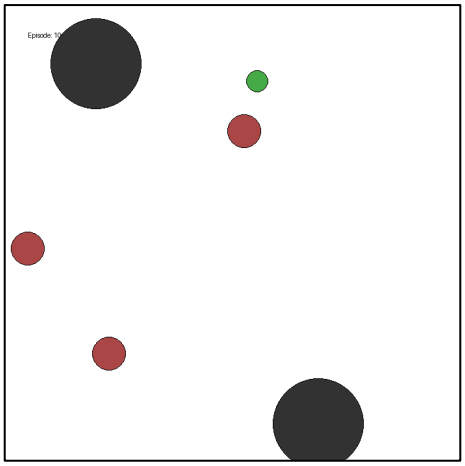


(7)　simple_tag_v3

エージェント数：4

良性エージェント：黄緑色
敵性エージェント：赤色
障害物：灰色

赤色の敵性エージェントは、黄緑色の良性エージェントに衝突すれば、正の報酬が得られる。黄緑色の良性エージェントは、赤色の敵性エージェントに衝突されると負の報酬を与えられる。黄緑色の良性エージェントは、赤色の敵性エージェントよりも早く移動することができる。灰色の障害物は、良性エージェントも敵性エージェントも通過することができない。敵性エージェントが協調して、良性エージェントを追い込むような行動が学習される。

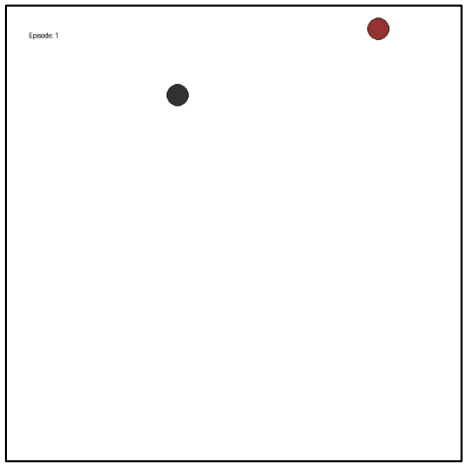


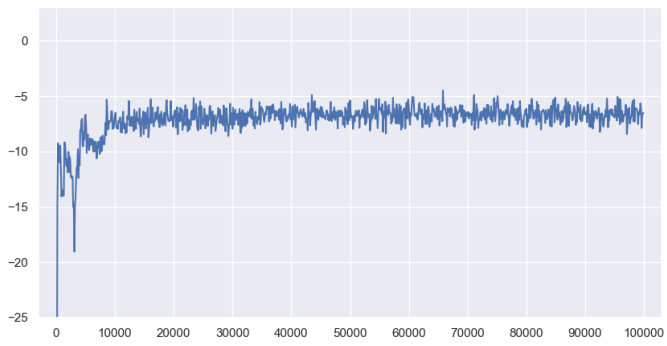
(8)　simple_v3

エージェント数：1

エージェント：赤色
ターゲット・ランドマーク：灰色

赤色のエージェントは、灰色のターゲット・ランドマークに近づくほど正の報酬が得られる。エージェントがターゲット・ランドマークに向かって移動するように行動が学習される。



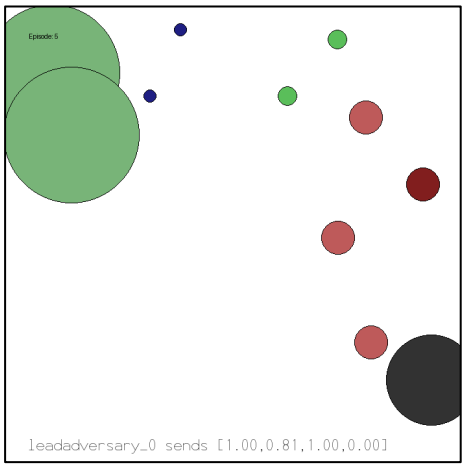


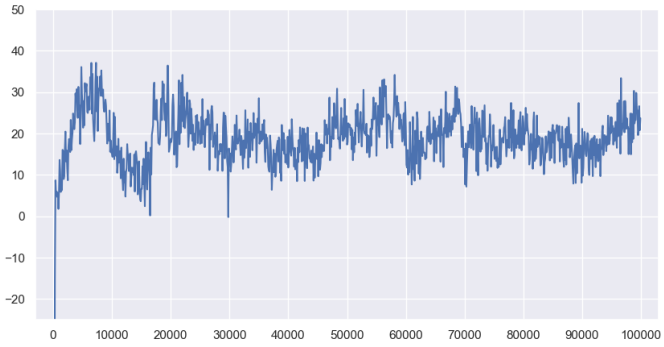
(9) simple_world_comm_v3

エージェント数:6

食べもの:青色
良性エージェント:黄緑色
敵性エージェントリーダー:濃赤色
敵性エージェントメンバー:薄赤色
障害物:灰色
森:緑色

良性エージェントは食べものを追い、食べものに近づくこと、または接触することで正の報酬を得られる。敵性エージェントは良性エージェントを追い、良性エージェントに近づくことまたは接触することで正の報酬を得られる。このとき、良性エージェントの位置を把握できるのは敵性エージェントリーダーだけであり、敵性エージェントリーダーは、敵性エージェントメンバーに情報を送信することで追跡指示を出す。環境にはエージェント以外に、森と障害物がある。森の内部にエージェントが侵入したとき、その位置が他のエージェントから観測できなくなる。障害物は各エージェントが通り抜けることができない。複合的な報酬関係での追跡動作が学習される。





5. まとめ

今回の取り組みで、AgileRL と PettingZoo を利用したマルチエージェント強化学習の基礎的な環境を構築することができた。

今後、下記のような項目に対して、研究を発展させたい。

- PettingZooMPE 環境をベースにして、オリジナルのマルチエージェント環境を設定し、挙動や学習過程を比較する。
- 良性エージェントと敵性エージェントの学習状態を、それぞれ別々の評価指標で測定する。

参考文献

[1] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, Igor Mordatch: Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environment, 2017.

[2] Johannes Ackermann, Volker Gabler, Takayuki Osa, Masashi Sugiyama: Reducing Overestimation Bias in Multi-Agent Domains Using Double Centralized Critics, 2019.

[3] Kai Liu, Yuyang Zhao, Gang Wang, Bei Peng: SA-MATD3:Self-attention-based multi-agent continuous control method in cooperative environments, 2021.

[4] Ustaran-Anderegg, N., & Pratt, M. AgileRL [Computer software]. https://github.com/AgileRL/AgileRL

[5] Terry, J., Black, B., Grammel, N., Jayakumar, M., Hari, A., Sullivan, R., Santos, L., Perez, R., Horsch, C., Dieffendahl, C., Williams, N., & Lokesh, Y. PettingZoo: Gym for multi-agent reinforcement learning [Computer software]. https://github.com/Farama-Foundation/PettingZoo