敵対強化学習におけるゲームの複雑性の定量評価

大嶋 真理絵　中田 亨 (産総研NEC-産総研AI連携研究室)

背景と目的

◆背景

敵対強化学習は、頑健な方策を学習でき、学習速度も向上するという利点がある[3]。しかし、敵対的エージェントは主となるエージェントの学習効果を向上させるための外乱エージェントである研究が多く、エージェント同士が対等に敵対する状況に関する研究は多くない。

◆目的

1. 敵対するエージェント同士による学習過程や結果の変化
2. 非対称なルールによる違い

これらを行動選択の不確かさ等の指標から考察

敵対強化学習とは

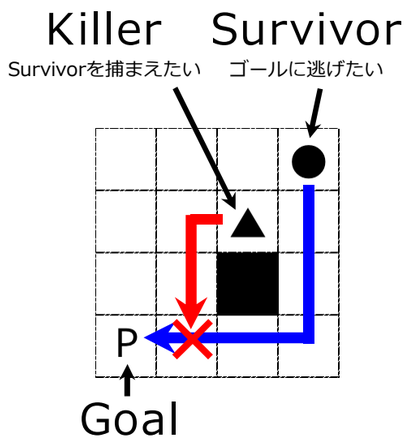
- 利害の一致しないエージェントを同一環境におき、それぞれ独立して強化学習をさせること
- 頑健な方策を学習でき、学習速度も向上する[3]

アプローチ

- 鬼ごっこゲームを敵対強化学習させ、行動選択の不確かさを測る
- ゲームにルールを追加して、行動選択の不確かさの差を測る

実験に用いるゲーム

◆基本ルール



勝敗
- SurvivorはGoalに到達したら勝ち
- KillerはSurvivorと同じ場所に到達したら勝ち
- タイムアウト（上限ステップ到達時）はKillerの勝ち

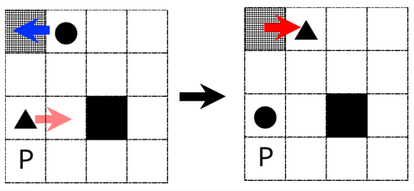
マップ
- 4x4のグリッドワールド
- 初期位置、障害物やゴールの位置は固定
- 障害物のある場所は通行できない

行動
- 行動選択肢は上下左右、その場にとどまる(Stay)
- 壁や障害物へ向かう行動は選択させない

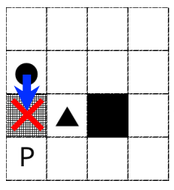
同時手番ゲームであり、不完全情報ゲーム

◆追加ルール

survivorによる位置交換（Exchange）
- 特定の場所に来ると、 両者の位置が入れ替わる
- Survivorのみ発動できる



Killerが罠を設置できる（Trap）
- Killerは1回だけ、Stayの代わりに罠を設置できる
- Survivorが罠にかかったらKillerの勝ち



罠の位置は双方わからない
Killerは罠を持っているかどうかはわかる

◆報酬

	Killer	Survivor
捕獲／ゴールによる勝利	500	
タイムアウトによる勝利	20	
行動ミス	-10	-
ターンコスト	1	-1

敗北時の報酬は勝利報酬の負値

行動ミス：
罠を使い切っているのに罠を設置しようとする

行動選択の不確かさ

$$\left(\begin{array}{c}\text{一定エピソードごとの}\\ \text{行動選択の不確かさの総和}\end{array}\right) = \sum_{s_i \in \text{全状態}} \left(f(s_i) \sum_{a_j \in \text{全行動}} \left(P(a_j|s_i) \log_2 \frac{1}{P(a_j|s_i)} \right) \right)$$

f(s):状態sへの遷移回数 P(a|s): 状態s_iにおいて行動a_jを選択した割合

↓ ステップ数による正規化

$$\left(\begin{array}{c}\text{1ステップごとの}\\ \text{行動選択の不確かさの}\\ \text{平均値}\end{array}\right) = \sum_{s_i \in S} \left(f(s_i) \sum_{a_j \in A} \left(P(a_j|s_i) \log_2 \frac{1}{P(a_j|s_i)} \right) \right) \div \text{全ステップ数}$$

◆例

状態	行動 a_1	行動 a_2
s_1	1	1
s_2	0	5

2回 x 1 bit = 2 bit
5回 x 0 bit = 0 bit
} 合計して 2 bit → ステップ数で正規化して 2 bit ÷ 7回 = 約0.29 bit / step

実験仮説

1. 自分に有利なゲームのほうが行動選択の不確かさは少なくなる←有利さを生かした戦略（必勝法）を構成する手数は多くないため
2. 学習が進むと、各プレイヤーの行動選択の不確かさは近い値に均衡する←相手が限られた手しか打ってこなければ、応答する側も少ない種類の手で対応すれば済むため
3. ルールの効果は線形に重畳する？？
4. 新戦略が発見されると、発見した方の行動選択の不確かさは減少し、相手は増加する

実験アルゴリズム

◆シングルQ学習

$$Q(s_t, at) \leftarrow Q(s_t, at) + \alpha[\gamma \max_a Q(s_{t+1}, a) - Q(s, a)]$$

s_t:現在の状態 a:選択行動 s_{t+1}:遷移後の状態 r:報酬 γ: 割引率 α: 学習率

◆状態表現

KillerとSurvivorの座標の組み合わせ（4次元、225状態）
Trapルールの場合、Killerは罠の残数を加えた5次元, 450状態

◆学習ターン

- Killerから学習を開始し、交互に学習させる
- Killerの初回学習ターンのみSurvivorの行動選択はStayのみ

学習パラメータ

割引率	0.98
学習率	1.0
ε	1.0~0.05
学習ターン数（プレイヤーあたり）	50,000
エピソード数（ターンあたり）	10
上限ステップ数（エピソードあたり）	30

実験結果

仮説1 ×

TrapはKiller有利だが、Killerの行動選択の不確かさは増加した

仮説2 △

Plain, Exchangeは45度線に近い値に収束するが、Trapを含むルールはさほど45度線に近づかない

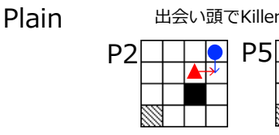
仮説3 ×

Plain→Trap+ExchangeはPlain→TrapとPlain→Exchangeの線形和にならない（収束点の関係）

仮説4 ○

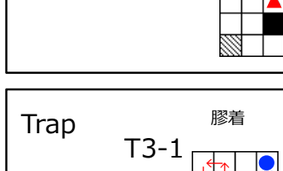
ExchangeのE5ではKillerが負けていたが、E7からタイムアウトに持ち込みE11から負けなくなる

収束点の関係

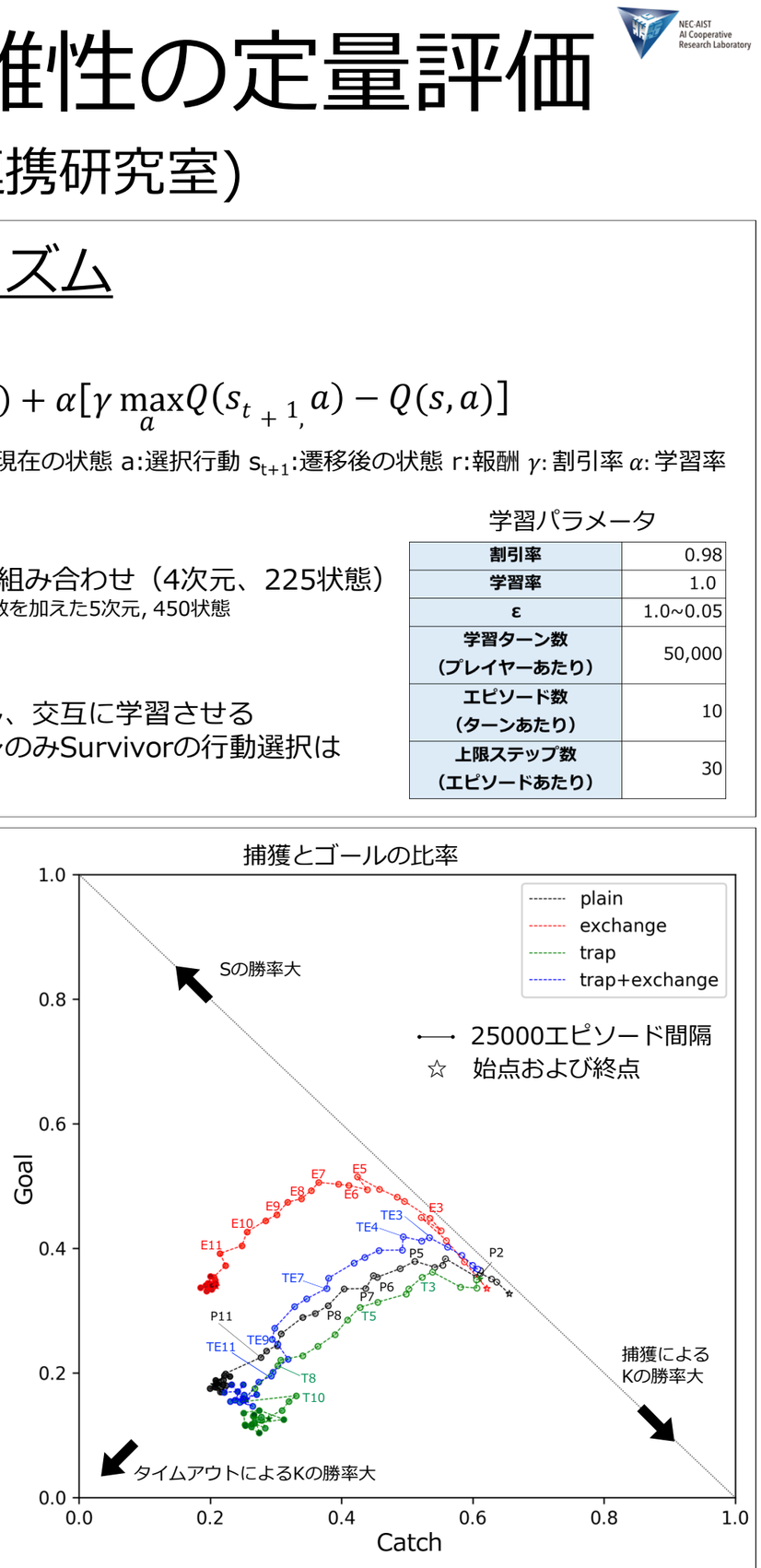


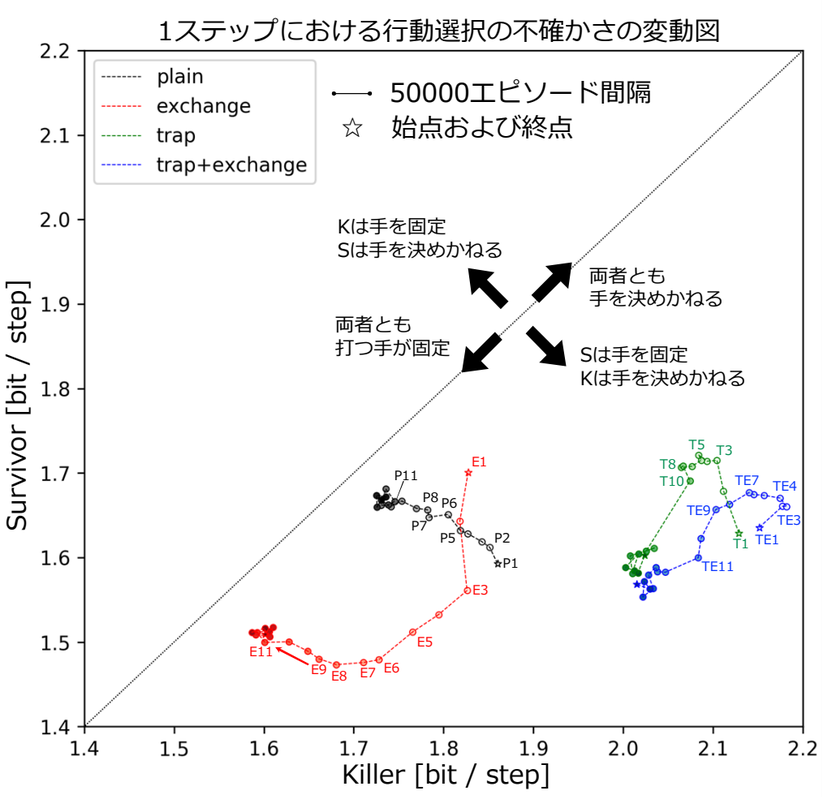
ルールの影響の重畳が線形だった場合のT+E

軌跡の形状

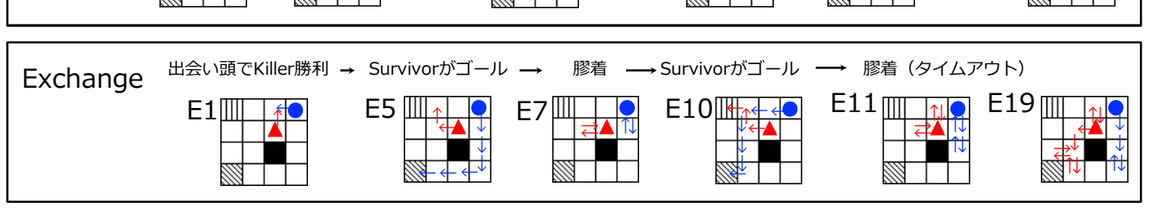


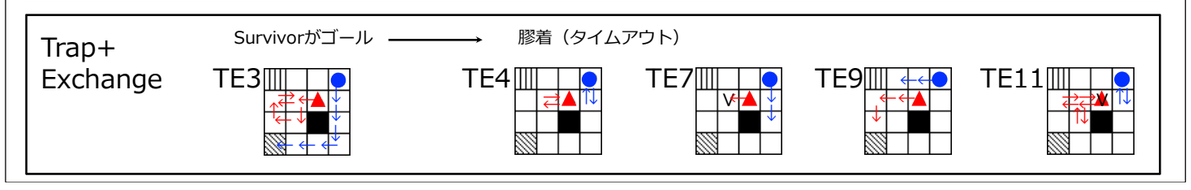
軌跡が上に凸のルール：Trap, T+E
Sが先に手を広めて学習し、そのあとKが手を広めて学習する
軌跡が下に凸のルール：Exchange
Sが先に打つ手を絞っていき、そのあとKが手を絞っていく
軌跡が直線のルール：Plain
Kは手を絞っていくが、Sは手を決めかねる





新戦略発見の例





まとめ

- 有利不利と行動選択の不確かさは単純な相関はしなかった
- 複数のルールの重畳は、片方を打ち消すこともある
- 行動選択の不確かさの変動図は、新戦略発見のタイミングを観察できた
- 行動選択の不確かさの変動図は、敵対強化学習の観察ツールになる

参考文献

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al.: “Human-level control through deep reinforcement learning”, Nature 518, pp.529-533, 2015.
[2] J. Heinrich, D. Silver: “Deep Reinforcement Learning from Self-Play in Imperfect-Information Games”, arXiv:1603.01121, 2016.
[3] L. Pinto, J. Davidson, R. Sukthankar, A. Gupta: “Robust adversarial reinforcement learning”, arXiv preprint arXiv:1703.02702, 2017.