分散的ベイズ推論による創発コミュニケーションに基づくマルチエージェント強化学習

江原広人 [1]　中村友昭 [1]　谷口彰 [2]　谷口忠大 [2]
[1] 電気通信大学　[2] 立命館大学
(h_ebara, nakamura)@radish.ee.uec.ac.jp
(a.taniguchi, taniguchi)@em.ci.ritsumei.ac.jp

概要

人は共同タスクを遂行する中で，他者や環境とインタラクションすることによって，協調行動を学習することができる．また，互いの内部状態を表現したメッセージを伝達し合うことで，互いの状態を知ることができる．このように，コミュニケーションを通じて互いに理解できる記号が創発される過程のことを，創発コミュニケーションと呼ぶ．これまで，マルチエージェント強化学習によって協調行動を学習する研究が行われおり，これらの研究では主にエージェント同士をネットワークで接続して学習するアプローチが取られていた．しかしこれは，独立した個体同士の創発コミュニケーションの観点からすると，不自然な設定である．そこで本稿では，創発コミュニケーションとマルチエージェント強化学習を組み合わせることで，エージェント間をネットワークで接続することなく，協調行動を学習できるモデルを提案する．実験では，エージェント同士が協調し，異なるゴールに到達する移動タスクを行い，適切な協調行動を学習できることを示す．また提案手法が，エージェント同士をネットワークで接続して学習する従来手法と同等の性能を発揮できることを示す．

1　はじめに

人間は他者や環境とのインタラクションによって，協調行動を学習することができる．また，他者の内部状態を表現したメッセージを通じて他者の状態を知ることにより，適切な行動を選択することができる．このようなコミュニケーションを通じて互いに理解できる記号が創発する過程は創発コミュニケーションと呼ばれており，記号や言語の形成を説明する創発コミュニケーションの構成論の探求は計算言語学において重要な課題である．近年では，そのような創発コミュニケーションとマルチエージェント強化学習を組み合わせた研究が行われている [1–5]．これらの研究では，コミュニケーションを通じて，複数のエージェントが協調行動を学習している．しかし，従来手法の多くは中央集権 (Centralized Training: CT) 型の学習である．CT 型のモデルは，エージェント同士がネットワークを介して接続されており，学習時に他者の内部状態や観測から計算された誤差情報が各ネットワークに逆伝播することで，各エージェントは他者の状態を考慮した行動を学習できる．一方で，人間は独立した個体であり，他者の内部状態から計算された情報を直接利用することはできない．すなわち，独立した個体同士の創発コミュニケーションの観点から考えると，CT 型は不自然な設定だといえる．したがって，各エージェントが自身の内部状態のみからメッセージを創発し，独立して学習を行う分散（Decentralized Training: DT）型の学習が適切だと考えられる．

一方，谷口は，記号は集合的な確率推論によって形成されているとする集合的予測符号化仮説 [6, 7] を提唱し，記号の推論手法としてメトロポリス・ヘイスティングス名付けゲーム (MHNG) [8, 9] を提案した．MHNG は確率的生成モデルによって記号創発の過程を定式化し，独立したエージェント間で，共有される記号を創発できる手法である．さらに，MHNG によって，特定の記号創発タスクにおける人の行動も説明できることが示されている [10]．

これまでに我々は，MHNG と強化学習を組み合わせることで，DT 型で協調行動を学習する手法を提案した [11]．しかし，この従来手法では離散的な状態・行動を用いていたため，限られたタスクのみにしか適用することができなかった．そこで，本稿では，MHNG と深層強化学習を組み合わせることで，連続的な状態行動空間で協調行動を学習できる DT 型のモデル MASAC-ECo (Multi-agent Soft Actor-Critic with Emergent Communication）を提案する．実験では，2 体のエージェントが創発されたメッセージを

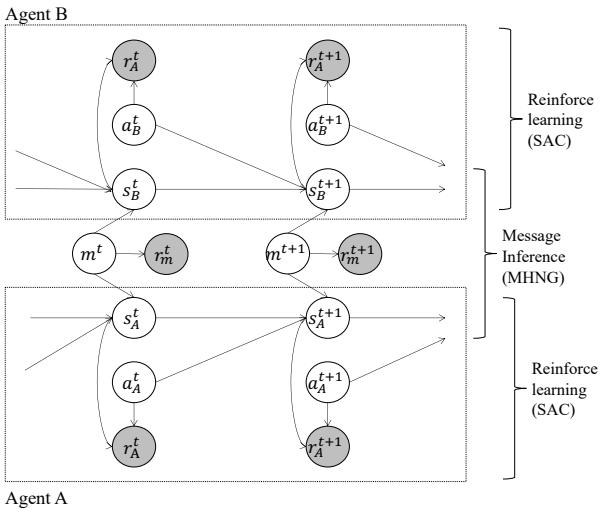

図 1 提案手法のグラフィカルモデル．各エージェントの状態・行動モデルである 2 つのマルコフ決定過程が，メッセージ m^t で接続された確率的生成モデルである．

表 1 各変数の詳細

s_A^t, s_B^t	各エージェントの状態
a_A^t, a_B^t	各エージェントの行動
r_A^t, r_B^t	各エージェントの報酬
r_m^t	協調行動の報酬
m^t	エージェント間でやり取りされるメッセージ

介してコミュニケーションすることで，協調行動の学習が可能であることを示す．また，CT 型の従来手法との性能を比較し，提案手法がそれらと同等の性能を発揮できることを示す．

2 提案手法

図 1 が，本稿で想定する確率的生成モデルのグラフィカルモデルである．各変数の詳細は表 1 の通りである．このモデルは，各エージェントの状態・行動モデルがマルコフ決定過程 (MDP) になっており，これらがメッセージ m^t で接続されたモデルである．m^t によって，各エージェントは他者の状態を間接的に知ることができるため，協調行動の生成が可能である．しかし，DT 型の学習でこれら全てのパラメータを推論することは困難である．そこで，本稿ではメッセージの推論部分と，MDP を学習する強化学習部分に分割し，メッセージの推論には MHNG を，強化学習には soft actor-critic (SAC) [12] を用いる．

2.1 MHNG によるメッセージの創発

2 体のエージェントが協調行動するためには，互いの状態を伝達するための記号が必要となる．ここで，図 1 のように，m^t という潜在変数から互いの状態 s_A^t, s_B^t と協調行動の報酬 r_m^t が生成されると仮定する．このモデルでは，m^t は互いの状態の決定に影響を与えるため，この潜在変数 m^t はメッセージと考えることができる．メッセージの創発は，各ステップのエージェントの状態 s_A^t, s_B^t と協調行動の報酬 r_m^t からメッセージ m^t を推論することに相当する．

$$m^t \sim p(\cdot|s_A^t, s_B^t, r_m^t) \tag{1}$$

しかし，式 (1) は自身からは観測できない相手の状態が含まれており，直接計算することができない．そこで，文献 [8,9] と同様に MHNG を用いる．MHNG では，一方のエージェントの提案分布からサンプリングした m^{t*} を相手に提案し，相手はそれを受理または棄却することを繰り返すことで，目標分布からのサンプルの生成が可能な手法である．まず，求めたいサンプルは両者の状態 s_A^t, s_B^t と協調行動の報酬 r_m^t の関係を表現したメッセージ m^t であるため，目標分布は次式となる．

$$P(m^t) = p(m^t|s_A^t, s_B^t, r_m^t) \tag{2}$$

$$\approx p(m^t|s_A^t, r_{m^t}^t)p(m^t|s_B^t, r_m^t) \tag{3}$$

式 (3) への変形には，Product of Experts (PoE) 近似を用いている．エージェント A がメッセージを提案する場合，提案分布は次式となる．

$$Q(m^{t*}|m^t) = p(m^{t*}|s_A^t, r_m^t) \tag{4}$$

式 (4) に従って，エージェント A は新たなサンプルを生成し，B に提案する．B は提案された m^{t*} を自身の予測に基づき，次式の受理確率 ψ に従って受理または棄却する．

$$\psi = \frac{P(m^{t*})Q(m^t|m^{t*})}{P(m^t)Q(m^{t*}|m^t)} \tag{5}$$

$$= \frac{p(m^{t*}|s_A^t, r_m^t)p(m^{t*}|s_B^t, r_m^t)p(m^t|s_A^t, r_m^t)}{p(m^t|s_A^t, r_m^t)p(m^t|s_B^t, r_m^t)p(m^{t*}|s_A^t, r_m^t)}$$

$$= \frac{p(m^{t*}|s_B^t, r_m^t)}{p(m^t|s_B^t, r_m^t)} \tag{6}$$

式 (5) は式 (3) と式 (4) により，式 (6) のように変形できる．式 (6) より，エージェント A から提案された m^{t*} の受理確率は，エージェント B のパラメータのみから計算することができる．つまり，相手の状態を直接観測することなく，メッセージの受理/棄却を判断することができる．

以上の手順を役割を交代しながら繰り返し，最適なメッセージを推論する．このメッセージのやり取りをするコミュニケーションによって，2 体のエージェントは他者の状態を間接的に知ることができ，

両者の状態に応じた最適な行動を選択することができる．

2.2 状態 s とメッセージ m に基づいた行動決定

各エージェント $i \in \{A, B\}$ は，自身の状態 s_i^t と 2.1 節で推論したメッセージ m^t に基づき，次式から行動 a_i^t を選択する．

$$a_i^t \sim \pi_i(\cdot|s_i^t, m^t) \tag{7}$$

π_i は各エージェントの方策である．各エージェントはメッセージによって間接的に相手の状態を知ることができるため，自身と相手の状態に応じた行動を選択することができる．つまり，メッセージには両者の行動を変化させる役割があり，コミュニケーションによって協調を促すように両者の行動を調整できる．また，選択した行動 a_i^t に対する価値 v_i^t は，次式の行動価値関数 Q_i から算出される．

$$v_i^t = Q_i(s_i^t, a_i^t, m^t) \tag{8}$$

π_i と Q_i をニューラルネットワークで近似した soft actor-critic を用いて，これらの関数を学習する．ネットワークのパラメータ φ_i，θ_i を用いて各エージェントの方策を π_{φ_i}，行動価値関数を Q_{θ_i} と表すと，最小化する目的関数は次式となる．

$$J_\pi(\varphi_i) = E\left[\alpha \log \pi_{\varphi_i}(a_i^t|s_i^t, m^t) - Q_{\theta_i}(s_i^t, a_i^t, m^t)\right] \tag{9}$$

$$J_Q(\theta_i) = E\left[\left(Q_{\theta_i}(s_i^t, a_i^t, m^t) - \hat{Q}(s_i^t, a_i^t, m^t)\right)^2\right] \tag{10}$$

ただし

$$\hat{Q}(s_i^t, a_i^t, m^t) = r_i^t + \beta r_m^t + \gamma E\left[V(s_i^{t+1}, a_i^{t+1}, m^{t+1})\right] \tag{11}$$

$$V(s_i^t, a_i^t, m^t) = E\left[Q_{\theta_i}(s_i^t, a_i^t, m^t) - \alpha \log \pi_{\varphi_i}(a_i^t|s_i^t, m^t)\right] \tag{12}$$

である．また，α はエントロピー正則化への重み，β は協調行動の報酬の重み，γ は割引率である．

3 実験

提案手法によって 2 体のエージェントが協調行動を学習できるかを検証するため，複数のエージェントが互いに協調し合い，それぞれ異なるゴールに到達することを目的とする移動タスクを行った．

3.1 実験設定

実験環境として，図 2 に示す Cooperative Navigation タスク [13] [1] を使用した．図 2 中の赤・青丸は各エージェント，黒点はゴールを表す．エージェントの状態 s は，2 次元の位置座標 $s = [x, y]$ である．行動 a は，5 次元のベクトル $a = [a_1, a_2, a_3, a_4, a_5]$ であり，各次元はそれぞれ進行方向 [止, 左, 右, 上, 下] の

1） https://github.com/openai/multiagent-particle-envs

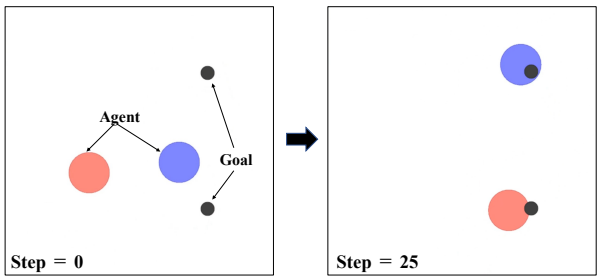

図 2 Cooperative Navigation のタスク環境

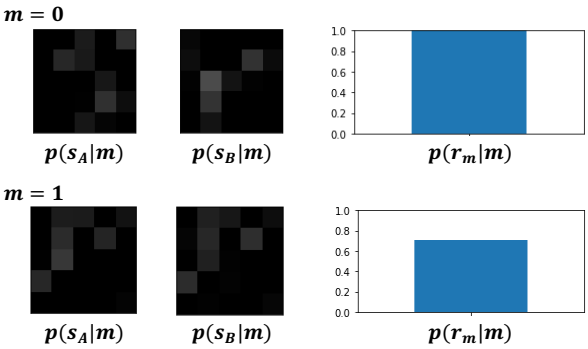

図 3 生成されたメッセージの例

重みを表している．報酬 r は，各エージェントに近い方のゴールとの距離 d の負の値 $-d$ とした．協調行動の報酬 r_m は，両者が衝突した場合は -1 とし，衝突していなければ 0 とした．メッセージ m は，推論されたメッセージの要素を 1 とした 64 次元のワンホットベクトルとした．

本実験では，従来手法である MADDPG [2]，BiCNet [14]，CommNet [15] と提案手法の性能を比較した．また，提案手法からコミュニケーションを省いたモデルと比較することで，創発されたメッセージが学習に与える影響について検証した．従来手法は，全て CT 型であり，全エージェントの内部状態や観測情報が共有される．また，他者からの誤差情報が自身のネットワークに伝播することで，他者の状態を知ることができる．これに対し，提案手法は DT 型であり，エージェント同士がネットワークを介して接続されていない．そこで提案手法では，両者の状態を表現するメッセージを創発し，そのメッセージを介してコミュニケーションすることで協調行動を学習する．

3.2 メッセージの創発

エージェントとゴールをランダムに環境内に配置し，移動させることで得られたデータを用いて，メッセージを学習した．1 エピソードを 25 ステップとし，1 エピソードが終了したら環境をリセットして，再び各エージェントとゴールをランダムな位置

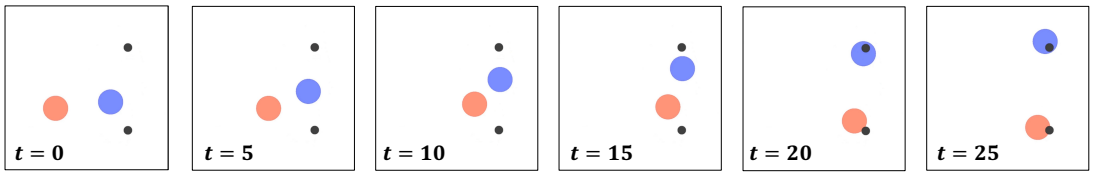

図 4　生成された軌道の例

表 2　各手法の平均累積報酬と総衝突回数 ($^*p < 0.05$)

	平均累積報酬	総衝突回数	学習方式
MASAC-ECo	**-41.82±0.99**	**57.84 ± 8.16**	DT
MASAC	-48.36 ±1.72*	100.0 ± 12.83	DT
MADDPG [2]	**-40.43 ± 0.70**	**55.28 ± 14.5**	CT
BiCNet [14]	-49.90 ± 2.59*	151.68. ± 18.49	CT
CommNet [15]	-46.27 ± 0.99*	58.12 ± 7.08	CT

に配置した．この工程を 1,200 エピソード分繰り返し，計 30,000 個の学習データ $d_i^t = [s_i^t, a_i^t, s_i^{t+1}, r_i^t, r_m^t]$ を取得した．次に，取得したデータから，互いの状態を表現するメッセージ m^t を MHNG によって推論した．本実験で扱う MHNG のモデルは，状態が離散値であることを仮定している．そのため，状態空間を 5 × 5 のグリッドに区切って番号を振り，各エージェントが存在するグリッド番号を状態 s_i^t として使用した．学習されたメッセージの例を図 3 に示す．図 3 中の $p(s_A|m), p(s_B|m)$ は，各エージェントが存在する確率分布を表す．また，$p(r_m|m)$ は，エージェントが高い報酬 $r_m = 0$ を得られる確率を表す．よって，$m = 0$ のメッセージは，互いが異なる位置に存在し，高い報酬が得られることを表現している．対して，$m = 1$ のメッセージは，互いが同じ位置に存在し，低い報酬が得られることを表現している．このことから，コミュニケーションによって互いの状態を適切に表現するメッセージを創発できたといえる．

3.3　協調行動の学習と性能評価

3.2 で得られた学習データ d_i^t を用いて，オフラインで各モデルのパラメータを学習した．提案手法の学習時のパラメータは，付録 4.1 に記載した通りである．提案手法の各エージェント $i \in \{A, B\}$ の方策は，学習データ d_i^t とメッセージ m^t を用いて，互いに独立して学習した．MASAC ではメッセージを必要としないため，各エージェント i の方策は d_i^t のみを用いて学習した．従来手法は他者の状態・行動が必要となるため，各エージェントの方策は両者のデータ d_A^t，d_B^t を同時に使用して学習した．

各手法の性能を評価するため，学習済みモデルのパラメータを固定して追加で 200 エピソード実行し，エピソード毎の平均累積報酬と総衝突回数を算出した．実験結果を表 2 に示す．表 2 中の「*」は，各手法と MASAC-ECo の累積報酬に対して t 検定を行った結果，有意水準 5%で有意差が認められたことを表している．MASAC-ECo では，相手の状態をメッセージを通して間接的に把握するため，CT 型の従来手法よりも不利な条件となっている．それにもかかわらず，表 2 より，MASAC-ECo と MADDPG との間に統計的な有意差はなく，同等の性能となった．また，メッセージを使用しない MASAC と比較して，MASAC-ECo の方が累積報酬が高く，衝突回数が少ないことが分かる．これらのことから，提案手法によって，両者が協調行動のための有益なメッセージを創発できたといえる．

図 4 は，生成された軌道の例であり，青のエージェントが赤のエージェントに道を譲り，初期位置から遠い方のゴールへ到達していることが分かる．これは，メッセージによって相手の位置を把握し，協調を優先する行動へと変化したためだと考えられる．このことから，創発されたメッセージを介してコミュニケーションし，適切な協調行動を学習できたことが示された．

4　おわりに

本稿では，創発コミュニケーションと深層強化学習を組み合わせることで，協調行動を学習するモデル MASAC-ECo を提案した．実験では，提案手法は DT 型であり，相手の状態を観測できない不利な条件であったが，相手の状態を直接観測できる CT 型の MADDPG と同等の性能を発揮した．この結果から，提案手法によって，協調行動のためメッセージを創発し，協調行動を学習できることを確認した．しかし，現状の課題として，MHNG によるメッセージの推論はオフラインでしか行えないという点がある．そのため，今回の実験ではオフライン学習で性能評価を行った．今後は，この問題を解決するために，オンライン学習ができるように提案モデルを拡張することを考えている．また，創発コミュニケーションの観点から，創発されたメッセージの持つ意味について分析することを考えている．

謝辞

本研究は，JST ムーンショット型研究開発事業 JPMJMS2011，JSPS 科研費 JP21H04904, JP23H04835 の支援を受けたものである．

参考文献

[1] Jakob N. Foerster, Yannis M. Assael, Nando de Freitas, Shimon Whiteson, “Learning to Communicate with Deep Multi-agent Reinforcement Learning”, Advances in Neural Information Processing Systems, Vol.29, 2016.

[2] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch, “Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments”, Conference on Neural Information Processing Systems, 2017.

[3] Jiechuan Jiang, Zongqing Lu, “Learning Attentional Communication for Multi-Agent Cooperation”, arXiv: 1805.07733, 2018.

[4] Angeliki Lazaridou, Marco Baroni, “Emergent Multi-Agent Communication in the Deep Learning Era”, arXiv:2006.02419, 2020.

[5] Rahma Chaabouni et al., “Emergent Communication at Scale”, International Conference on Learning Representations, 2022.

[6] 谷口忠大，“集合的予測符号化仮説 - 記号創発ロボティクスと言語進化の新展開に向けて”，人工知能学会全国大会，4H3-OS-6b-01，2023

[7] Tadahiro Taniguchi, “Collective Predictive Coding Hypothesis: Symbol Emergence as Decentralized Bayesian Inference”, PsyArXiv preprint, doi: https://doi.org/10.31234/osf.io/d2ty6, 2023

[8] Tadahiro Taniguch et al., “Emergent Communication through Metropolis-Hastings Naming Game with Deep Generative Models”, Advanced Robotics, Vol. 37, Issue 19, pp. 1266-1282, 2023

[9] Yoshinobu Hagiwara, Kazuma Furukawa, Akira Taniguchi, Tadahiro Taniguchi, “Multiagent Multimodal Categorization for Symbol Emergence: Emergent Communication via Interpersonal Cross-modal Inference”, Advanced Robotics, Vol. 36, Issue 5-6, pp. 239-260, 2022.

[10] Ryota Okumura, Tadahiro Taniguchi, Yoshinobu Hagiwara and Akira Taniguchi, “Metropolis-Hastings algorithm in joint-attention naming game: Experimental semiotics study”, Frontiers in Artificial Intelligence, vol. 6, 2023

[11] Tomoaki Nakamura, Akira Taniguchi, Tadahiro Taniguchi, “Control as Probabilistic Inference as an Emergent Communication Mechanism in Multi-Agent Reinforcement Learning”, arXiv, 2307.05004, 2023

[12] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, Sergey Levine, “Soft Actor-Critic: Off-policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor”, International Conference on Machine Learning, 2018.

[13] Igor Mordatch, Pieter Abbeel, “Emergence of Grounded Compositional Language in Multi-Agent Populations”, arXiv:1703.04908, 2017.

[14] Peng Peng et al. “Multiagent Bidirectionally-Coordinated Nets”, arXiv:1703.10069, 2017.

[15] Sainbayar Sukhbaatar, Arthur Szlam, Rob Fergus, “Learning Multiagent Communication with Backpropagation”, Conference on Neural Information Processing Systems, 2016.

4.1 付録

実験 3.3 での提案手法の学習時のパラメータ詳細を以下の表 3 に示す.

表 3 提案手法の学習時のパラメータ

パラメータ	値
エントロピー正則化項の重み (α)	0.5
協調行動の報酬の重み (β)	3.0
割引率 (γ)	0.95
バッチサイズ	1024
各ネットワークの中間層の数	3
各ネットワークのユニット数	64
学習率	0.01