分散的ベイズ推論としての マルチエージェント強化学習と記号創発

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

人間は会話や文章などの言語 (メッセージ) を通じて他者とコミュニケーションを取ることで，互いに協調した行動を学習することができる．ロボット同士が人間と同様のアプローチで協調行動を学習するためには，ロボット間で互いに理解できる共通言語を創発するモデルが必要となる．本稿では，エージェント自らが創発した記号を用いてコミュニケーションすることで，協調行動を学習・生成することができる確率的生成モデルを提案する．実験では，学習したモデルを実環境のタスクに適用し，2 台のロボットが創発した記号によるコミュニケーションを通して，環境に適した協調行動を生成できることを示す．

1　はじめに

人間は生まれてから未分化な認識世界の中で活動を始め，外部の環境との相互作用を通じて様々な概念や言語，行動を獲得することができる．また，獲得した言語を用いて他者とコミュニケーションを取ることで，互いに協調した行動を学習することができる．このように，人間同士が記号を介してコミュニケーションを取ることで，両者の間で理解できる共通言語が形成されていく過程を，コミュニケーションの創発と呼ぶ．これまでに，コミュニケーションの創発に基づくアプローチで，エージェント同士で記号を創発し，タスクを実行する様々な研究がなされてきた [1, 2]．さらに，近年ではコミュニケーションの創発を深層強化学習に拡張することで，より複雑なタスクへの応用が可能であることが示されている [3, 4, 5, 6]．一方で，谷口らによって，人間の言語獲得の基礎となる一部の発達過程を実現したメトロポリス・ヘイスティングス名付けゲーム

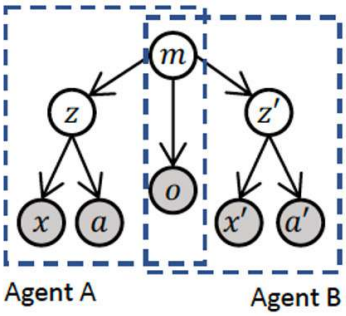

図 1　提案手法のグラフィカルモデル

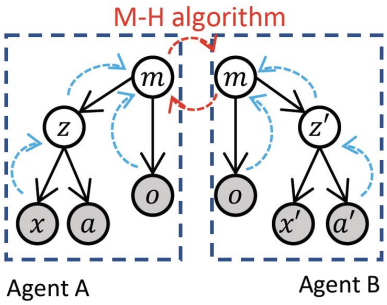

図 2　M-H 法により分割したモデル

と呼ばれる言語ゲームが提案されている [7, 8]．文献 [7, 8] では，同ゲームを用いて 2 体のエージェントがコミュニケーションを取ることで記号を創発し，物体の概念を形成できることを示している．この手法では，確率的生成モデルを用いることでコミュニケーションの創発を実現している．しかし，このアプローチで協調行動を学習し，実環境のタスクを遂行できるようなモデルは，未だに確立されていない．そこで本稿では，メトロポリス・ヘイスティングス名付けゲームと強化学習と組み合わせることで，生成モデル的アプローチでエージェントが協調行動を学習することができるモデルを提案する．実験では，学習したモデルを実環境のタスクに適用し，2 台のロボットが創発した記号を用いて環境に適した協調行動を選択できるかを検証した．

2 提案手法

2.1 モデルと生成過程

図 1 に提案手法のグラフィカルモデルを示す．z はエージェントの内部状態を表現する潜在変数であり，m はエージェント間で生成されるメッセージ（共通言語）を表現する潜在変数である．提案モデルでは，2 体のエージェント間で共有されたメッセージ m によって各エージェントの内部状態 z, z' を決定し，協調行動を生成することを仮定している．

$$z \sim P(z|m) \tag{1}$$

$$o \sim P(o|m) \tag{2}$$

$$a \sim P(a|z) \tag{3}$$

$$x \sim P(x|z) \tag{4}$$

x, a は各エージェントの観測と行動である．本稿では，確率的な推論 (Control as Inference [9]) による制御に基づいた，両エージェントの行動の最適制御問題を考える．そのため，一般の強化学習における報酬の代わりに，最適性変数と呼ばれる二値の確率変数 $o \in \{0, 1\}$ を用いる．最適性変数とは，両エージェントの行動の適切さを評価する変数であり，$o = 1$ が最適であることを表し，行動の最適性を確率 $p(o = 1|m)$ で表現する．

2.2 メトロポリス・ヘイスティングス名付けゲームによる m の推論

内部状態 z, z' と最適性 o から，各エージェントの行動パターンと，それらに対する最適性の関係を表現するメッセージ m を推論する．

$$m \sim p(m|z, z', o) \tag{5}$$

しかし，式 (5) では，自身からは観測できない相手の内部状態 z' 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 [7, 8] と同様に，メトロポリス・ヘイスティングス名付けゲームを用いる．この手法では Metropolis-Hastings (M-H) 法 [10] を用いることでモデルを図 2 のように分割することができ，互いに内部状態を知ることなくメッセージ m を推論することができる．M-H 法では，両者は提案分布からサンプリングした m^* 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を生成することが可能な手法である．まず，求めたいサンプルは両者の内部状態 z, z' と最適性変数 o の関係を表現したメッセージ m であるため，目標分布は次式となる．

$$P(m) = p(m|z, z', o) \tag{6}$$

$$\approx p(m|z, o)p(m|z', o) \tag{7}$$

ただし，この式変形には Product-of-Experts (PoE) 近似を用いた．次に，エージェント A がメッセージを提案するため，提案分布は次式となる．

$$Q(m^*|m) = p(m|z, o) \tag{8}$$

この提案分布従い，エージェント A が新たなサンプル m^* を生成し，B に提案する．B は提案された m^* を自身の予測に基づき，次式の受理確率に従って，受理または棄却するかを決定する．

$$r = \frac{P(m^*)Q(m|m^*)}{P(m)Q(m^*|m)} \tag{9}$$

$$= \frac{p(m^*|z, o)p(m^*|z', o)p(m|z, o)}{p(m|z, o)p(m|z', o)p(m^*|z, o)} \tag{10}$$

$$= \frac{p(m^*|z', o)}{p(m|z', o)} \tag{11}$$

式 (9) は式 (7) と式 (8) を用いることで，式 (11) のように変形することができ，エージェント A から提案された m^* の受理確率は，エージェント B のパラメータのみで計算できる．すなわち，相手の内部状態を知ることなく，M-H 法に基づく受理/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，収束するまで繰り返し，最適なメッセージ m を推論する．本稿では，このメッセージ m をやり取りすることがコミュニケーションであると考え，このコミュニケーションによってマルチエージェントの最適な行動選択が可能となる．

2.3 エージェントの内部状態 z の推論

2.2 節で学習したメッセージと各エージェントの観測に基づき，次式のようにエージェントの内部状態 z のパラメータを推論する．

$$z \sim p(z|m, a, x) = \frac{p(x, a|z)p(z|m)}{p(x)p(a)} \tag{12}$$

また，行動予測時にはクロスモーダル推論によって行動決定する．

$$a \sim p(a|m, x) = \int_z p(a|z)\frac{p(x|z)p(z|m)}{p(x)}dz \tag{13}$$

本稿では，このような潜在変数の推論とクロスモーダル推論が可能な Multimodal Latent Dirichlet Allocation(MLDA)[11] を拡張した手法 [12] を用いる．

Algorithm 1 Infference of m, z, z' by M-H naming game

function M-H(z, z', o, m)
  $m^* \sim p(m^*|z, o)$
  $r = \min(1, \frac{p(m^*|z',o)}{p(m|z',o)})$
  $u \sim$ Uniform(0,1)
  if $u \leq r$ **then**
    return m^*
  else
    return m
  end if
end function

for $t = 1$ to T **do**
  // Infference of inner state of each agent
  $z \sim p(z|m, a, x)$
  $z' \sim p(z'|m, a', x')$

  // Determine the message through communication
  for $n = 1$ to N **do**
    // Agent A proposes to Agent B
    $m \leftarrow$ M-H(z, z', o, m)
    // Agent B proposes to Agent A
    $m \leftarrow$ M-H(z', z, o, m)
  end for
end for

2.4 メッセージの創発と行動決定

学習時には 2.2 節のコミュニケーションによるメッセージの推論と，2.3 節の各エージェントの内部状態の推論を Algorithm 1 のように繰り返すことで，各確率分布のパラメータを更新する．これにより，各エージェントの観測 x, x' と最適性 o から協調行動を表現するメッセージ m を創発することができる．

行動決定時には，学習したパラメータを固定し最適性を $o = 1$ として Algorithm 1 を実行し，メッセージ m と内部状態 z, z' を推論する．推論されたパラメータに基づいて，式 (13) のクロスモーダル推論に基づき，行動決定することができる．

3 実験

実ロボットを用いたタスクにより，提案手法の有効性を確認した．

3.1 実験設定

実験環境を図 3 に示す．ロボットは ROBOTIS 社の turtlebot3(手前側) と turtlebot2(奥側) を使用した[1)]．障害物 (大) の位置は固定，障害物 (小) は左右どちらかに配置し，両ロボットが衝突せずに障害物を回避するルートの選択が学習可能か検証した．両ロボットの頭部には，現在の位置から障害物までの距離を計測できるレーザーレンジファインダ (LRF) が搭載されており，これを用いて自己位置を推定し，障害物を回避して移動することができる．ここで問題となるのが，turtlebot2 の LRF の位置が障害物 (小) よりも高い位置にあり，障害物 (小) を検知できないという点である．これにより，turtlebot2 は障害物 (小) が置かれた方のルートを選択した場合，回避できずに障害物 (小) と衝突する可能性がある．そこで，提案手法を用いてロボット間でコミュニケーションを取ることで，障害物の位置に応じて最適な行動を選択できるよう，メッセージ m を推論する必要がある．

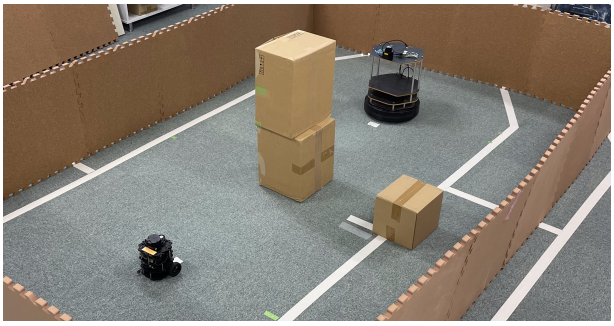

図 3 実験環境

観測 x は正面から左右斜め 45° までの障害物までの距離を 1° ずつ計測した長さ 90 のベクトルとし，行動 a は左または右回りの 2 パターンとした．また，最適性 o は，互いが干渉せずに障害物を回避する行動を選択すれば $o = 1$，それ以外は $o = 0$ とした．本実験でのタスクでは，両エージェントは障害物の位置に対して計 8 パターンの行動が考えられる．よって，m は 8 次元の多項分布とし，各パターンの行動ペアが 1 つのクラスに分類されるようにした．

以降，turtlebot3 をエージェント A，turtlebot2 をエージェント B とする．

3.2 学習

数パターンの観測 x，行動 a，最適性 o のデータを計 40 個を生成し，相互更新の回数 $T = 20$ としてモデルを学習した．学習されたメッセージ m と最適性 o の関係を表 1 に示す．表より，最適性に基づきメッセージが学習されていることが分かる．$m = 1$ では，エージェント A から見て左側に障害物が置かれ，A と B がそれぞれ左回りで移動する行動が表現

1) https://www.turtlebot.com

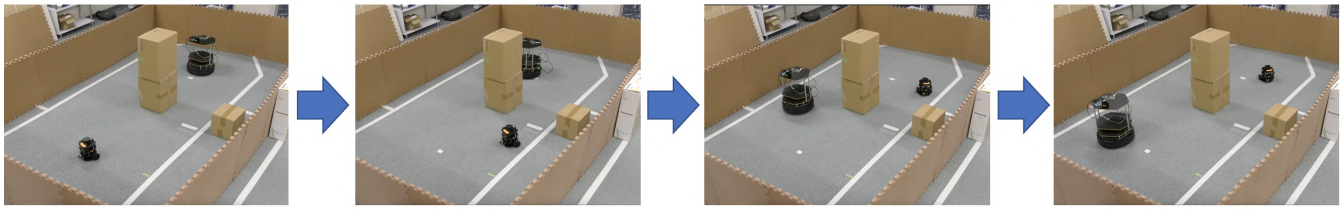

図 4　予測時の行動の様子

表 1　メッセージ m と最適性 o の関係

m	$o=0$	$o=1$
0	0.986	0.014
1	0.083	0.917
2	0.981	0.019
3	0.981	0.019
4	0.976	0.024
5	0.976	0.024
6	0.981	0.019
7	0.020	0.980

したメッセージであった．$m=7$ では，エージェント A から見て右側に障害物が置かれ，A と B がそれぞれ右回りで移動する行動が表現したメッセージであった．このように，環境・行動・最適性を表現したメッセージを学習できていることが確認できた．

3.3　最適行動の予測

学習済みのモデルを用いて，障害物 (小) を左右ランダム置いた環境で，両エージェントが提案手法を用いてコミュニケーションすることで，最適な行動を予測できるか検証した．予測の際の相互更新の回数は $T=5$ とし，障害物の位置をランダムに変更し 20 回試行した．両エージェントの行動の一例を図 4 に，各障害物配置における行動の成功回数と生成されたメッセージを表 2 に示す．この表より，適切なメッセージを生成することで，障害物を回避し，衝突しない適切なルートを選択できていることが分かる．

次に，相互更新回数 T，すなわちコミュニケーション繰り返す回数を変化させ，M-H 法によるコミュニケーションによるメッセージの生成の有効性を確認した．図 5 が各相互更新回数 T における最適行動を選択できた確率である．図より，コミュニケーションの回数が増えるにつれて，最適な行動選択ができるようになっていることが分かる．エージェント A は障害物をどちらも知覚できるため，コミュニケーションしなくとも高い確率で最適行動を選択できている．一方，エージェント B は障害物 (小) を検知できないため，メッセージが内部状態の推定に反映されない $T=1$ では，自身の観測だけでは最適な行動を選択することはできていない．その後，T が増加するに従い，メッセージによってエージェント B の欠損している障害物 (小) の情報を補うことができ，最適な行動を選択できている．

表 2　予測に基づく行動の成功回数と生成されたメッセージ

	障害物 (小)：左	障害物 (小)：右
成功回数 (回)	20/20	20/20
生成された m	1	7

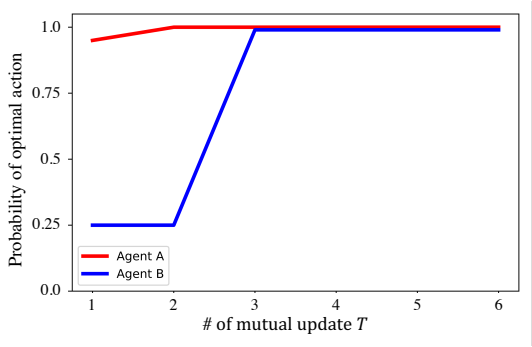

図 5　相互回数 T における最適行動の選択確率

4　おわりに

本稿では，エージェント間でコミュニケーションを取ることで，協調行動を学習・生成できる確率的生成モデルを提案した．実験では，提案手法を用いて両ロボットがコミュニケーションを取り，適切なメッセージを学習できることを確認した．また，コミュニケーションにより，最適な行動を選択できることを確認した．しかし，本稿で扱ったタスクは，最適性変数が即時的に決定するワンショットな意思決定タスクであり，累積報酬を最大化する必要があるタスクにはそのままでは適用できない．そこで，本手法を時間発展させ，より複雑なロボットタスクへと応用することを今後の課題とする．

謝辞

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

参考文献

[1] Diane Bouchacourt and Marco Baroni, “MissTools and Mr Fruit: Emergent communication in agents learning about object affordances”, in Proceedings of the 57th annual meeting of the association for computational linguistics (ACL), pp. 3909–3918, 2019

[2] Katrina Evtimova, Andrew Drozdov, Douwe Kiela, and Kyunghyun Cho, “Emergent Communication in a Multi-Modal, Multi-Step Referential Game”, International Conference on Learning Representations, 2018

[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] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch, “Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments”, arXiv: 1706.02275, 2020

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

[7] Tadarhito Taniguch et al., “Emergent Communication through Metropolis-Hastings Naming Game with Deep Generative Models”, arXiv: 2205.12392, 2022.

[8] 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

[9] Sergey Levine, “Reinforcement learning and control as probabilistic inference: Tutorial and review.” arXiv preprint arXiv:1805.00909, 2018

[10] Hastings, W. K., “Monte Carlo sampling methods using Markov chains and their applications”, Biometrika 57, pp. 97-109, 1970

[11] Tomoaki Nakamura, Takayuki Nagai, and Naoto Iwahashi, “Grounding of Word Meanings in Multimodal Concepts Using LDA”, IROS2009, pp.3943-3948, 2009

[12] Tomoaki Nakamura, Takayuki Nagai, and Tadahiro Taniguchi, “Serket: An architecture for connecting stochastic models to realize a large-scale cognitive model”, Frontiers in Neurorobotics, vol.12, pp.1-16, 2018.