3R-2

通信規約学習の拡張による協調精度の向上

葛西　達也　　小林　隼人　　篠原　歩
東北大学　大学院情報科学研究科

1　はじめに

自律エージェント集団の協調問題において，協調精度の向上を促進させる有益な方法として，エージェント間通信がある．しかし，適切な通信規約を定義するためには，扱う問題に対する正しい知識が不可欠であり，未知問題においては設計者の知識の有無によって通信の効果が左右される．そこで，我々は，エージェント自身に自律的に通信規約を学習させるための方法の構築・実現を目指している．

先行研究として，葛西らはマルチエージェント強化学習 (Multi-Agent Reinforcement Learning : MARL) の枠組みにおいて，通信方策と行動方策を同時に学習させる *Signal Learning* [1] を提案している．本論文では，従来の signal learning を拡張し，不完全知覚環境下における例題に対して拡張手法の実装を行う．そして，従来手法と比べ拡張手法の方を用いた方が，学習精度が向上する事を示す．さらに，学習を通して，本来欠落している情報を，通信内容に補完させるような学習が行われる事で，決定的な最適方策を獲得可能である事を示す．

2　関連研究

MARL では，S を環境の状態集合，A をエージェントの行動集合とした時，各々のエージェントが協調的な方策 $\pi : S \rightarrow A$ を学習する．通信を用いる場合には，事前に適切な通信規約を構築する必要がある．しかし，未知問題においては，効果的な情報や通信規約を定義するのは困難である．

葛西らは，MARL の枠組みにおいて，自律的に通信規約を学習させる事を目的とした，*Signal Learning* (SL) [1] を提案している．SL では，明示的に意味を付与されていないメッセージ集合を M とした場合に，エージェントが通信方策 $\pi_c : S \rightarrow M$ と行動方策 $\pi_a : S \times M \rightarrow A$ を同時に学習する事で，通信規約の自律学習を行っている．葛西らは，通信無しの場合と比べ，SL を用いた方が精度が向上する他，$|M|$ を増加させた場合に，同時に精度も向上する事を実験的に示している．この事は，SL の学習過程において，問題解決に有効な意味がメッセージに込められたという事である．故に SL は，未知問題に対して通信を用いる場合に非常に有益である．通信を用いる学習に関する研究はいくつか報告されているが [2] [3]，意味を持たないメッセージを扱う場合の議論はなされていない．

3　Signal Learning の拡張

拡張は，通信方策を $\pi_c : S \rightarrow M$ から $\pi_c : S \times M \rightarrow M$ へと変更するシンプルなものである．SL では，エージェントは観測した状態のみに基づいて，送信メッセージを決定する．それに対して SLM では，観測した状態と，受信メッセージの両方に基づいて，送信メッセージを決定する．故に，我々はこの拡張手法を *SL with Messages* (SLM) と呼ぶ．Algorithm 1 は，二体のエージェントの MARL における，各エージェントの 1 ステップサイクルを示している．なお，本論文では，実験の初段階として二体のエージェントの場合の MARL に対して焦点を当てる．

Algorithm 1 各エージェントの 1 ステップサイクル

1: 環境から状態 $s \in S$ を観測
2: 他エージェントからのメッセージ $m \in M$ を受信
3: 環境に対して行動 $a = \pi_a(s, m)$ を実行
4: 他エージェントへメッセージ $m' = \pi_c(s, m)$ を送信
5: 環境から報酬 $r \in \boldsymbol{R}$ を観測
6: 報酬 r に基づいて方策 π_c と π_a を更新

4　例題

我々は，SL と SLM の性能の違いを明らかにするため，図 1 で示すような例題に取り組む．例題は，各エージェントが状態 Start/Goal (SG) をスタートし，用意されたボタンを ON にした後，両エージェントが状態 SG まで帰還した場合にゴールとなる．ボタンを ON にするためには，両エージェントが一度は同時に状態 Button (B) に滞在する必要がある．また，その他の状態を C と表記する．各エージェントは自分の状態，すなわち S = {SG, C, B}, しか観測する事ができない．取りうる行動は前進と後退，すなわち A = {Fore, Back} である．

本例題では，各エージェントは，他エージェントの状態を観測できない事に加え，ボタンの ON/OFF を記憶する事ができない．故に，通信を用いない場合，極めて学習が困難であると同時に，SL では仕組み上，決定的な最適方策の獲得が期待できない．この事は実験の考察部において詳しく述べる．

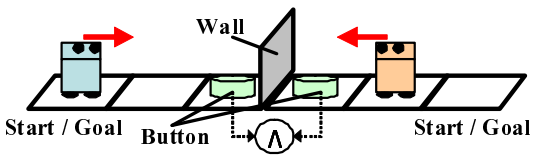

図 1: 例題

Improvement of the Performance in the Learning of Communication Codes using Received Messages
Tatsuya KASAI, Hayato KOBAYASHI and Ayumi SHINOHARA
Graduate School of Information Sciences, Tohoku University

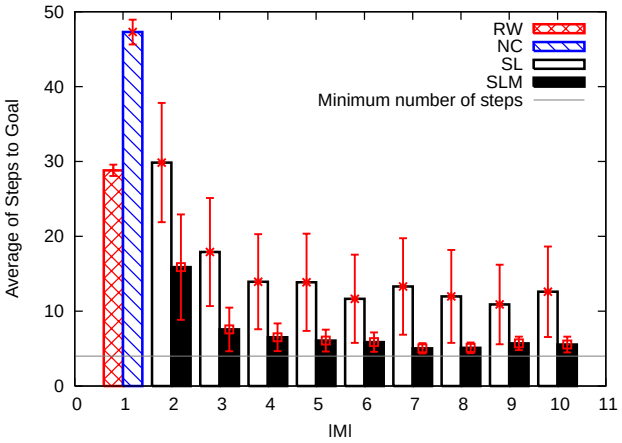

図 2: RW，NC，SL，SLM の比較

5 実験

$|M|$ を 2〜10 まで変化させた場合の，SL と SLM の比較実験を行う．$|M|$=1 の場合は，SL と SLM は等価であり，この場合は通信無し (NC) と表記する．さらに，本例題の難しさを示すための指標として，各ステップでランダムに行動を選択する Random Walk (RW) も加え，SL，SLM，NC，RW の 4 つに対して実験を行った．

強化学習アルゴリズムには，non-MDPs において頑健である Profit Sharing (PS)，行動選択手法にはルーレット選択をそれぞれ採用した．PS における Q 値の更新は，1 エピソード終了時に，エピソード中の各 $s_t \in S$，$a_t \in A$ に対して，$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + f(t, r, T)$ の更新式を用いて一括で行われる．T はエピソードの終端時刻，r は T のみで獲得される報酬，f は報酬割当関数である．実験では，$f(t, r, T) = r \cdot \gamma^{T-t-1} / \log(t)$ を用い，割引率 γ=0.5，r=100 とした．100 ステップ以内にゴールに到達できない場合，そのエピソードは無効とし，Q 値の更新を行わずにリスタートさせた．

評価は，10,000 エピソードを 1 試行とした場合に，終端 100 エピソード分のゴール到達ステップ数の平均値で行う．図 2 は，各 $|M|$ における，100 試行の平均を示したものであり，エラーバーは標準偏差を表す．

6 評価と考察

NC と SL を比較した場合，SL が極めて良い結果を示している事がわかる．これは，SL では，学習を通して何らかの有益な意味をメッセージに発生させているからである．つまり，SL では $\pi_c : S \rightarrow M$ によって，自分の状態をメッセージに込める事が可能であり，互いの位置を知る事ができる．それに対し NC では，互いの位置すら知る術がない．また，NC と SL は仕組み上，いずれもボタンの ON/OFF を保持する事はできない．RW よりも NC の方が悪い結果であるのは，通信を用いない一般的な方法では，この例題の学習が難しい事を実証している．SLM と SL の比較では，SLM が明らかに SL より頑健な結果を示している事がわかる．これは，SLM が，SL よりも有益な情報をメッセージに発生させているからと考えられる．SLM において，

表 1: 成功試行の割合

$\|M\|$	2	3	4	5	6	7	8	9	10
SL (%)	0	1	1	0	0	0	0	0	2
SLM (%)	31	34	42	47	41	45	60	54	48

表 2: 決定的最適方策 (M={1,2})

$S \times M$	(SG,1)	(SG,2)	(C,1)	(C,2)	(B,1)	(B,2)
π_a	Fore	Fore	Fore	Back	Back	Back
π_c	1	1	1	2	2	2

メッセージに込められたであろうより有益な情報として考えられるのは，ボタンの ON/OFF の情報であると考えられる．

本例題の場合，ボタンの ON/OFF の情報を保持出来なければ，決定的最適方策は存在しない．従って，決定的最適方策の有無に注目する事で，SLM がボタンの ON/OFF の情報を保持しているのかどうかがわかる．決定的最適方策が存在し，学習によってそれを獲得できた場合，例題は最短 (=4) ステップでゴールする事ができる．そこで，1 試行の学習が終了した後，獲得された方策のテストを行った．その結果，4 ステップでゴール到達に至った試行を成功試行として，100 試行中の成功試行の割合を示したものが表 1 である．表 1 より，SLM には明らかに成功試行の割合が高く，決定的最適方策が存在しうる事がわかる．事実，SLM では決定的最適方策が存在し，学習によってその獲得が行われる．表 2 は，SLM に存在する決定的最適方策の内，実際に獲得された最もシンプルな例 ($|M|$=2) である．表 2 では，$\pi_c : S \times M \rightarrow M$ において，$1 \in M$ を OFF，$2 \in M$ を ON として保持している事がわかる．

SLM では本来観測不可能な情報を，学習過程において，メッセージに補完する能力が備わっている事がわかる．このプロセスは，不完全知覚状態の分離に相当する．従って，部分観測可能マルコフ決定過程 (Partially Observable Markov Decision Processes : POMDPs) における環境下において，SLM が有効に働く可能性があるが，より具体的な検証が必要である．

7 まとめ

本論文では，SL を拡張した SLM を用いる事で，顕著に協調精度の向上が成される事を示した．また，SL には存在しない決定的最適方策に関して，SLM では獲得可能である事を実験的に示した．

参考文献

[1] T. Kasai, H. Tenmoto, and A. Kamiya, “Learning of Communication Codes in Multi-Agent Reinforcement Learning Problem,” in *Proc. of the 2008 IEEE Conference on Soft Computing in Industrial Applications (SMCia/08)*, 2008, pp. 1–6.

[2] C. V. Goldman and S. Zilbserstein, “Decentralized Control of Cooperative Systems: Categorization and Complexity Analysis,” *Journal of Artificial Intelligence Research*, vol. 22, pp. 143–174, 2004.

[3] D. Chakraborty and S. Sen, “Computing effective communication policies in multiagent systems,” in *AAMAS’07*, 2007, pp. 153–155.