The Size of Message Set Needed for the Optimal Communication Policy

Tatsuya Kasai, Hayato Kobayashi, and Ayumi Shinohara

Graduate School of Information Sciences, Tohoku University, Japan
kasai@shino.ecei.tohoku.ac.jp, kobayashi@shino.ecei.tohoku.ac.jp,
ayumi@ecei.tohoku.ac.jp

Abstract. Communication is a key for facilitating multi-agent coordination on cooperative problems. In our previous work, we proposed *Signal Learning* (SL) and *Signal Learning with Messages* (SLM) by which agents learn local policies of communication and action simultaneously in Multi-Agent Reinforcement Learning (MARL) framework. Our experimental results showed that both SL and SLM can improve the performance of agents' coordination.
In this paper, we focus on theoretical analysis of the conditions for constructing optimal local policies on SL and SLM framework in Decentralized Partially Observable Markov Decision Processes with Communication (Dec-POMDP-Com) models. As main results, we obtain the minimum required sizes of the message set for off-line computation of optimal local policies on SL and SLM. In addition, we report experimental results indicating that the extra messages make some positive effect in learning processes when the size of the message set is larger than the minimum required size based on theoretical analysis.

1 Introduction

In Multi-Agent Reinforcement Learning (MARL), each agent i learns a cooperative policy $\delta_i^A : \Omega_i \rightarrow A_i$, where Ω_i and A_i are a set of observations and a set of actions of agent i, respectively. If we aim to utilize communication to facilitate multi-agent coordination, we must construct communication codes so that agents can communicate with each other. However, it is a hard task since we usually do not know workable communication codes and/or information on unknown problems.

In our previous work [1], we proposed *Signal Learning* (SL), which allows agents to learn communication codes autonomously. By using SL, agents can learn local communication policy ($\delta_i^M : \Omega_i \rightarrow M_i$) and local action policy ($\delta_i^A : \Omega_i \times \boldsymbol{M}_i^{recv} \rightarrow A_i$) simultaneously in MARL framework, where M_i is a set of messages of agent i, whose meanings are not predetermined explicitly, e.g., $M_i = \{1, 2, 3\}$, and $\boldsymbol{M}_i^{recv}$ represents a set of joint messages that agent i receives from the other agents. We experimentally showed that the performance becomes better as the size $|M_i|$ of the message set increases. It should be noted that the messages in M_i have no meaning in the initial phase of learning. Our results

suggest that some beneficial meaning can emerge through learning process in SL, and SL is extremely helpful for unknown problems.

We recently proposed an extended version of SL, *Signal Learning with Messages* (SLM) [2]. The extension is simply the change of communication policy from $(\delta_i^M : \Omega_i \to M_i)$ to $(\delta_i^M : \Omega_i \times \boldsymbol{M}_i^{recv} \to M_i)$. Although the change seems to be surficial, our experiments showed that the performance of SLM is clearly better and more robust than that of SL on a simple task. Surprisingly, the simple task was a good example to prove that SLM has the ability to acquire a deterministic optimal policy, which cannot be achieved by SL.

As far as we know, our studies are rare ones to utilize meaningless messages to enhance multi-agent coordination in MARL framework. Although there are several related studies about learning of communication policy in MARL framework [3–9], these studies predetermine the meanings of messages. Jim and Giles [10, 11] also utilize meaningless messages to enhance multi-agent coordination, while they adopt Genetic Algorithms (GA) for learning of communication codes. They reported about the evolution of languages in their experiments, which is similar to our results. One of the common shortcomings of our studies and their studies is lack of theoretical justification, which would be very useful for designers to construct an efficient multi-agent system.

In this paper, we focus on shifting the direction of our studies about learning of communication codes to theoretical aspects. We recall the Decentralized Partially Observable Markov Decision Process model with Communication (Dec-POMDP-Com) from Goldman and Zilberstein [12]. The Dec-POMDP-Com is a useful tools to analyze multi-agent system in a decision-theoretic context. There are several studies about communication policy on the Dec-POMDP-Com model [13–17], most of which focus on theoretical analysis, such as complexity analysis of computing optimal policy on various versions of the model. While they are mostly interested in off-line computation of the optimal policy with a history of past observations and received joint messages, we have been interested in on-line computation of a better policy only with the a last observation and received joint message, i.e., on SL and SLM. As a first step to theoretical aspects, we address theoretical analysis about off-line computation of the optimal policy on SL and SLM.

The rest of this paper is organized as follows. In Section 2, we explain the Dec-POMDP-Com model and optimal solutions in the model, and in Section 3, we formally define policies on SL and SLM and express the difference from the standard definition. We theoretically analyze the condition of the message set, i.e., its minimum required size, so that optimal policies of SL and SLM can achieve the value of the optimal policy on the standard definition in Section 4. In Section 5, we report experimental results indicating that the extra messages make some positive effect in learning processes when the size of the message set is larger than the minimum required size based on theoretical analysis. In Section 6, we describe our conclusions and future work.

2 Dec-POMDP-Com

Dec-POMDP-Com is a decision theoretic model proposed by Goldman and Zilberstein [12], which can handle a decentralized multi-agent system, where agents can communicate with each other. Pynadath and Tambe proposed a similar model, Communication in a Markov Team Decision Process (COM-MTDP) [18]. We chose Dec-POMDP-Com simply because we do not utilize belief states in this paper. We should be able to make the same arguments on COM-MTDP, since Seuken and Zilberstein [17] proved Dec-POMDP-Com and COM-MTDP are equivalent under the *perfect recall* assumption, i.e., if an agent has access to all of its received information. A Dec-POMDP-Com is defined as follows.

Definition 1 (Dec-POMDP-Com). *A Decentralized Partially Observable Markov Decision Process with Communication is given by the tuple*

$$DPC := \langle I, S, \boldsymbol{\Omega}, \boldsymbol{A}, \boldsymbol{M}, C, P, R, O, T \rangle,$$

where

- $I := \{1, \cdots, n\}$ *is a finite set of agents, indexed* $1, \cdots, n$.
- S *is a finite set of global states, with distinguished initial state* s_0.
- $\boldsymbol{\Omega} := \prod_{i \in I} \Omega_i$ *is a finite set of joint observations, which consists of a finite set* Ω_i *of observations for agent* i.
- $\boldsymbol{A} := \prod_{i \in I} A_i$ *is a finite set of joint actions, which consists of a finite set* A_i *of actions for agent* i.
- $\boldsymbol{M} := \prod_{i \in I} M_i$ *is a finite set of joint messages which consists of a finite set* M_i *of messages for agent* i. *We define a finite set of joint messages that agent* i *receives from the other agents by* $\boldsymbol{M}_i^{recv} := \prod_{j \in I - \{i\}} M_j$. $\epsilon_\sigma \in M_i$ *is the null communication, i.e., sending an empty message.*
- $C : \boldsymbol{M} \to \Re$ *is a cost function.* $C(\boldsymbol{m})$ *represents the total cost of transmitting the messages sent by all agents. The cost of* ϵ_σ *is 0. We assume that the* C *is a constant function throughout this paper, since we do not predetermine the meanings of messages.*
- $P : S \times \boldsymbol{A} \times S \to [0, 1]$ *is a transition probability function.* $P(s, \boldsymbol{a}, s') := p(s'|s, \boldsymbol{a})$ *represents the probability of moving from global state* $s \in S$ *to global state* $s' \in S$ *when the agents take joint action* $\boldsymbol{a} \in \boldsymbol{A}$.
- $R : S \times \boldsymbol{A} \times S \to \Re$ *is a reward function.* $R(s, \boldsymbol{a}, s')$ *represents the reward for executing joint action* $\boldsymbol{a} \in \boldsymbol{A}$ *in global state* $s \in S$, *resulting in global state* $s' \in S$.
- $O : S \times \boldsymbol{A} \times S \times \boldsymbol{\Omega} \to [0, 1]$ *is an observation probability function.* $O(s, \boldsymbol{a}, s', \boldsymbol{o}) := p(\boldsymbol{o}|s, \boldsymbol{a}, s')$ *represents the probability of receiving joint observation* $\boldsymbol{o} \in \boldsymbol{\Omega}$ *when the agents take joint action* $\boldsymbol{a} \in \boldsymbol{A}$ *in global state* $s \in S$, *resulting in global state* $s' \in S$.
- T *is a (possibly infinite) time horizon in which the agents take their actions.*

The interaction among the agents on Dec-POMDP-Com is described as the following process. In a state $s \in S$, each agent i sends a message $m \in M_i$ and

performs an action $a \in A_i$ according to its *local policy* accessing only the information that the agent possesses. After executing a joint action $\boldsymbol{a} \in \boldsymbol{A}$, which consists of the actions of all agents, the state s moves to a state $s' \in S$ according to the transition probability function $P(s, \boldsymbol{a}, s')$. Each agent i receives an observation $o \in \Omega_i$ according to the observation probability function $O(s, \boldsymbol{a}, s', \boldsymbol{o})$ and a reward $r = R(s, \boldsymbol{a}, s')$.

We formally define a *local policy* δ_i *for agent* i as a pair of a local action policy δ_i^A and a local communication policy δ_i^M defined below, i.e., $\delta_i := \langle \delta_i^A, \delta_i^M \rangle$.

Definition 2 (Local Action Policy). *A* local action policy δ_i^A for agent i *is a mapping from the set* ${\Omega_i}^*$ *of histories of observations, and the set* $(\boldsymbol{M}_i^{recv})^*$ *of histories of received joint messages, to the set* A_i *of actions. That is,*

$$\delta_i^A : {\Omega_i}^* \times (\boldsymbol{M}_i^{recv})^* \to A_i.$$

Definition 3 (Local Communication Policy). *A* local communication policy δ_i^M for agent i *is a mapping from the set* ${\Omega_i}^*$ *of histories of observations, and the set* $(\boldsymbol{M}_i^{recv})^*$ *of histories of received joint messages, to the set* M_i *of messages. That is,*

$$\delta_i^M : {\Omega_i}^* \times (\boldsymbol{M}_i^{recv})^* \to M_i.$$

We denote the sets of local action policies δ_i^A and local communication policies δ_i^M for agent i by D_i^A and D_i^M, respectively. The set of local policies δ_i for agent i is defined by $D_i := D_i^A \times D_i^M$. Let $\boldsymbol{D} := \prod_{i \in I} D_i$. We call a tuple $\boldsymbol{\delta} \in \boldsymbol{D}$ of the local policies for all agents, a *joint policy*.

Solving a Dec-POMDP-Com means finding an optimal joint policy that maximizes the expected total rewards in the Dec-POMDP-Com. The optimal joint policy is formalized as

$$\boldsymbol{\delta}^* := \arg\max_{\boldsymbol{\delta} \in \boldsymbol{D}} V_{\boldsymbol{\delta}}^T(s_0),$$

where $V_{\boldsymbol{\delta}}^T(s_0)$ represents the expected total rewards after following a joint policy $\boldsymbol{\delta}$ from initial state s_0 within T time steps. We call it *the value of the joint policy* and define as follows.

Definition 4 (Value of a Joint Policy). *The* value of a joint policy $\boldsymbol{\delta}$ in initial state s_0 for time horizon T *is given by*

$$V_{\boldsymbol{\delta}}^T(s_0) := \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \left(R(s_t, \boldsymbol{a}_t, s_{t+1}) - C(\boldsymbol{m}_t) \right) \,\middle|\, s_0, \boldsymbol{\delta} \right],$$

where $s_t \in S$, $\boldsymbol{a}_t \in \boldsymbol{A}$, *and* $\boldsymbol{m}_t \in \boldsymbol{M}^{recv}$ *represent a state, a joint action and a received joint message at time step* t, *respectively, and* $\gamma \in [0, 1)$ *is a discount factor.*

3 Learning of Communication Codes

We proposed *Signal Learning* (SL) [1] and *Signal Learning with Messages* (SLM) [2], which allows agents to learn communication codes autonomously using some reinforcement learning algorithm in the one step dynamics shown in Algorithm 1. In SL, each agent sends a message depending only on an observation, while in SLM, each agent sends depending on both an observation and a received message. We showed that the performance of SLM is better than that of SL, since much more information is available in SLM.

Algorithm 1 One-step dynamics of agent i in SL / SLM

1: perform the action $\delta_i^A(o_t, \boldsymbol{m}_t)$ in the environment
2: receive an observation $o_{t+1} \in \Omega_i$ from the environment
3: send the message $\delta_i^M(o_{t+1})$ / $\delta_i^M(o_{t+1}, \boldsymbol{m}_t)$ to the other agents
4: receive a joint message $\boldsymbol{m}_{t+1} \in \boldsymbol{M}_i^{recv}$ sent by the other agents
5: receive a reward r_{t+1} from the environment
6: update δ_i^A and δ_i^M based on the reward r_{t+1}

In this section, we formalize SL and SLM in decision theoretic context.

3.1 Signal Learning

A local policy on SL is defined as a pair of a local action policy and a communication policy defined as follows.

Definition 5 (Local Action Policy on SL). *A* local action policy δ_i^A for agent i on SL *is a mapping from the set Ω_i of observations, and the set $\boldsymbol{M}_i^{recv}$ of received joint messages, to the set A_i of actions. That is,*

$$\delta_i^A : \Omega_i \times \boldsymbol{M}_i^{recv} \to A_i.$$

Definition 6 (Local Communication Policy on SL). *A* local communication policy δ_i^M for agent i on SL *is a mapping from the set Ω_i of observations, to the set M_i of messages. That is,*

$$\delta_i^M : \Omega_i \to M_i.$$

While a local policy on the standard definition allows an agent to access all of the history of its own observations and received joint messages, a local action policy on SL only allows an agent to access the last observation and received joint message, and a local communication policy on SL only allows an agent to access the last observation. Note that the set of local policies on SL is a subset of D, the set of local policies on the standard definition. Thus, we can mostly use the same notation in the Section 2. We denote the set on SL corresponding to D_i^A, D_i^M, D_i, and $\boldsymbol{D}$ by $D_i^{A,SL}$, $D_i^{M,SL}$, D_i^{SL}, and $\boldsymbol{D}^{SL}$, respectively.

3.2 Signal Learning with Messages

SLM is an extension of SL, and the only difference is local communication policy. A local communication policy on SLM is defined as follows.

Definition 7 (Local Communication Policy on SLM). *A* local communication policy δ_i^M for agent i on SLM *is a mapping from the set Ω_i of observations, and the set $\boldsymbol{M}_i^{recv}$ of received joint messages, to the set A_i of actions. That is,*

$$\delta_i^M : \Omega_i \times \boldsymbol{M}_i^{recv} \to M_i.$$

In the same manner as SL, we use the notations of $D_i^{A,SLM}$, $D_i^{M,SLM}$, D_i^{SLM}, and $\boldsymbol{D}^{SLM}$. Since $\boldsymbol{D}^{SL} \subset \boldsymbol{D}^{SLM} \subset \boldsymbol{D}$, the following equation clearly holds for any Dec-POMDP-Com:

$$\max_{\boldsymbol{\delta} \in \boldsymbol{D}^{SL}} V_{\boldsymbol{\delta}}^T(s_0) \leq \max_{\boldsymbol{\delta} \in \boldsymbol{D}^{SLM}} V_{\boldsymbol{\delta}}^T(s_0) \leq \max_{\boldsymbol{\delta} \in \boldsymbol{D}} V_{\boldsymbol{\delta}}^T(s_0).$$

4 Theoretical Analysis

The main objective of this section is to clarify the minimum required size $|M_i|$ of the message set for each agent i so that optimal policy on SL and SLM can achieve the value of the optimal policy on the standard definition in a decision theoretic context.

We start to refer the next useful theorem proved by Goldman and Zilberstein.

Theorem 1 (Goldman and Zilberstein [13]). *Let $V_{\boldsymbol{\delta},\boldsymbol{M}}^T(s_0)$ be the value of a joint policy $\boldsymbol{\delta}$ with respect to a joint message set $\boldsymbol{M}$. For any Dec-MDP-Com with constant message cost, the value of the optimal joint policy with respect to any joint message set $\boldsymbol{M}$ is not greater than the value of the optimal joint policy with respect to the joint message set $\boldsymbol{M}' := \boldsymbol{\Omega}$. That is*

$$\forall \boldsymbol{M}, \quad \max_{\boldsymbol{\delta} \in \boldsymbol{D}} V_{\boldsymbol{\delta},\boldsymbol{M}}^T(s_0) \leq \max_{\boldsymbol{\delta} \in \boldsymbol{D}} V_{\boldsymbol{\delta},\boldsymbol{M}'}^T(s_0),$$

where T and s_0 are the time horizon and the initial state in the Dec-MDP-Com, respectively.

This theorem means that the optimal local communication policy of each agent is to send its own observation in a Dec-MDP-Com. A Dec-MDP-Com is a *jointly fully observable* Dec-POMDP-Com, where there exists a mapping $J : \boldsymbol{\Omega} \to S$ such that whenever $O(s, \boldsymbol{a}, s', \boldsymbol{o})$ is non-zero then $J(\boldsymbol{o}) = s'$. Since the jointly full observability means that the combination of the agents' observations leads to the global state, the theorem is intuitively acceptable. Note that the theorem assumes no delay of the system, that is, agents can instantaneously communicate with each other.

Goldman and Zilberstein also proved that the optimal joint policy does not need the histories of the past observations and received messages of each agent in Corollary 4 in their paper [13]. Theorem 1 and their corollary immediately

yield the following corollary, which clarifies the sufficient condition of the size $|M_i|$ of the message set for each agent i in a Dec-MDP-Com so that an optimal joint policy on SL can achieve the value of an optimal policy on the standard definition.

Corollary 1. *For any Dec-MDP-Com with constant message cost, if the size $|M_i|$ of the message set of each agent i satisfies the condition,*

$$|M_i| \geq |\Omega_i|, \tag{1}$$

then the value of the optimal joint policy on SL is equal to the value of the optimal joint policy on the standard definition. That is

$$\max_{\boldsymbol{\delta} \in \boldsymbol{D}^{SL}} V_{\boldsymbol{\delta}}^{T}(s_0) = \max_{\boldsymbol{\delta}' \in \boldsymbol{D}} V_{\boldsymbol{\delta}'}^{T}(s_0),$$

where T and s_0 are the time horizon and the initial state in the Dec-MDP-Com, respectively.

Note that the condition (1) in Corollary 1 is not a necessary condition such that there exists an optimal joint policy on SL that achieves the value of an optimal joint policy on the standard definition. This is because if a Dec-MDP-Com is *fully observable*, where each agent can always observe its current global state, there clearly exists such an optimal policy on SL without communication, i.e., $|\boldsymbol{M}| = 0$.

Next, let us define *deterministic* Dec-POMDP-Com instead of Dec-MDP-Com in order to analyze SLM.

Definition 8 (Dec-POMDP-Com with Deterministic Transitions). *We say that a Dec-POMDP-Com has* deterministic transitions*, if for any state $s \in S$ and any joint action $\boldsymbol{a} \in \boldsymbol{A}$, there exists a state $s' \in S$ such that $P(s, \boldsymbol{a}, s') = 1$.*

When a Dec-POMDP-Com has deterministic transitions, there exists a *transition mapping* $f^{trn} : S \times \boldsymbol{A} \to S$, such that $f^{trn}(s, \boldsymbol{a}) = s'$ if and only if $P(s, \boldsymbol{a}, s') = 1$. In this case, we can simplify the notation of its observation probability function $O(s, \boldsymbol{a}, s', \boldsymbol{o})$ by using a function $G(s', \boldsymbol{o})$ in the next lemma.

Lemma 1. *If a Dec-POMDP-Com has deterministic transitions, then there exists a function $G : S \times \boldsymbol{\Omega} \to [0, 1]$ such that $O(s, \boldsymbol{a}, s', \boldsymbol{o}) = G(s', \boldsymbol{o})$ for any states $s, s' \in S$, any joint action $\boldsymbol{a} \in \boldsymbol{A}$, and any joint observation $\boldsymbol{o} \in \boldsymbol{\Omega}$.*

Proof: Let $G(s', \boldsymbol{o}) := p(\boldsymbol{o}|s')$. Since the Dec-POMDP-Com has deterministic transitions, the transition probability $P(s, \boldsymbol{a}, s')$ is either 1 or 0. Therefore,

$$\begin{aligned} G(s', \boldsymbol{o}) &= \sum_{s \in S, \boldsymbol{a} \in \boldsymbol{A}} p(\boldsymbol{o}|s, \boldsymbol{a}, s') p(s, \boldsymbol{a}|s') \\ &= \sum_{s \in S, \boldsymbol{a} \in \boldsymbol{A}} O(s, \boldsymbol{a}, s', \boldsymbol{o}) P(s, \boldsymbol{a}, s') \\ &= O(s, \boldsymbol{a}, s', \boldsymbol{o}). \end{aligned}$$

■

Definition 9 (Deterministic Observable Dec-POMDP-Com). *We say that a Dec-POMDP-Com is* deterministic observable*, if for any states $s, s' \in S$ and any joint action $\boldsymbol{a} \in \boldsymbol{A}$, there exists a joint observation $\boldsymbol{o} \in \boldsymbol{\Omega}$ such that $O(s, \boldsymbol{a}, s', \boldsymbol{o}) = 1$.*

When a Dec-POMDP-Com is deterministic observable, there exists an *observation mapping* $f^{obs} : S \times \boldsymbol{A} \times S \to \boldsymbol{\Omega}$, such that $f^{obs}(s, \boldsymbol{a}, s') = \boldsymbol{o}$ if and only if $O(s, \boldsymbol{a}, s', \boldsymbol{o}) = 1$.

Definition 10 (Deterministic Dec-POMDP-Com). *We say that a Dec-POMDP-Com is* deterministic*, if it has deterministic transition and is deterministic observable.*

From Lemma 1, when the Dec-POMDP-Com is deterministic, we can denote the observation mapping by $f^{obs}(s')$ as a substitution for $f^{obs}(s, \boldsymbol{a}, s')$. In this case, we refer to the number of states corresponding to the observation $o \in \Omega_i$ of agent i by

$$S_i^{obs}(o) := \{s \in S \mid f_i^{obs}(s) = o\},$$

where $f_i^{obs} : S \to \Omega_i$ is the observation mapping of agent i.

The next theorem clarifies the sufficient condition of the size of the message set in a deterministic Dec-POMDP-Com so that an optimal joint policy on SLM can achieve the value of an optimal policy on the standard definition.

Theorem 2. *For any deterministic Dec-POMDP-Com with constant message cost, if the size $|M_i|$ of the message set of each agent i satisfies the condition,*

$$|M_i| \geq \max_{j \in I} \max_{o \in \Omega_j} |S_j^{obs}(o)|, \tag{2}$$

then the value of the optimal joint policy on SLM is equal to the value of any joint policy on the standard definition. That is

$$\max_{\boldsymbol{\delta} \in \boldsymbol{D}^{SLM}} V_{\boldsymbol{\delta}}^T(s_0) = \max_{\boldsymbol{\delta}' \in \boldsymbol{D}} V_{\boldsymbol{\delta}'}^T(s_0),$$

where T and s_0 are the time horizon and the initial state in the Dec-POMDP-Com, respectively.

Proof: We only prove the case of $n = 2$. Let $i, j \in \{1, 2\} : i \neq j$ be the indexes of one agent and the other agent, respectively.

Since $|M_j| \geq \max_{o \in \Omega_i} |S_i^{obs}(o)|$, agent i can decompose own observation $o \in \Omega_i$ to the current global state $s' \in S$ by using the message $m \in M_j$ received from agent j. That is, there exists a function $g_i : \Omega_i \times M_j \to S$ such that for any observation $o \in \Omega_i$ of agent i and any state $s \in S_i^{obs}(o)$ corresponding to the observation o, there exists a message $m \in M_j$ of agent j such that $g_i(o, m) = s$. By using the function g_i, agent j can always construct a communication policy $\delta_j^M \in D_j^{M,SLM}$ so that agent i can specify the global state. That is, for any joint action policy $(\delta_1^A, \delta_2^A) \in D^{A,SLM}$, there exists a communication policy

$\delta_j^M \in D_j^{M,SLM}$ of agent j such that for any joint observation $(o_1, o_2) \in \boldsymbol{\Omega}$ and any joint message $(m_1, m_2) \in \boldsymbol{M}$,

$$s' = g_i(f_i^{obs}(s'), \delta_j^M(f_j^{obs}(s'), m_i)),$$

where s' is the next global state, i.e., $s' := f^{trn}(s, (\delta_1^A(o_1, m_2), \delta_2^A(o_2, m_1)))$.

Since the optimal joint action only depends on the global state, there is no more effective communication policies than the above-mentioned communication policy $\delta_j^M \in D_j^{M,SLM}$ of agent j. For the same reason, the value of the optimal joint policy on SLM achieves the value of the optimal joint policy on the standard definition, even though a policy on SLM can only access the last observation and received message. ■

The condition (2) in Theorem 2 is not a necessary condition for the same reason as SL. Furthermore, we can easily construct a deterministic Dec-POMDP-Com such that for any index $i \in I$,

$$|M_i| < \max_{j \in I} \max_{o \in \Omega_j} |S_j^{obs}(o)|.$$

From Corollary 1, if a Dec-POMDP-Com is jointly fully observable, such a Dec-POMDP-Com is a concrete example such that the condition (2) in Theorem 2 does not hold, although the optimal joint policies on SLM need explicit communications, (i.e., $|\boldsymbol{M}| > 0$), to achieve the value of an optimal joint policy on the standard definition.

Let us define *Cooperative Button Pushing (CBP)* problem for the arguments in the sequel.

Definition 11 (Cooperative Button Pushing Problem). *We define Cooperative Button Pushing (CBP) Problem as follows (see Fig. 1):*

- *The goal of this problem is that two agents, starting from their own Start/Goal (SG) grids, go back to the SG grids after switching the status of buttons from* OFF *to* ON*, where agents repeat to move forward or backward.*
- *In order to switch the status of buttons to* ON*, agents must occupy their own Button (B) grids at the same time at least once.*
- *The other grids represent by Center (C).*
- *The agents receive a positive reward only when the agents reach the goal.*
- *The agents cannot observe the grids of other agent and the status of buttons (also cannot remember the status of buttons).*
- *The policy so that the agents can always reach the goal in minimum steps is optimal.*

The next fact shows that there exists a nontrivial Dec-POMDP-Com such that the condition (2) in Theorem 2 does not hold.

Fact 1 *There exists a deterministic Dec-POMDP-Com with constant message cost, which is not jointly fully observable, such that*

$$\max_{\boldsymbol{\delta} \in \boldsymbol{D}^{SLM}} V_{\boldsymbol{\delta}}^T(s_0) = \max_{\boldsymbol{\delta}' \in \boldsymbol{D}} V_{\boldsymbol{\delta}'}^T(s_0),$$

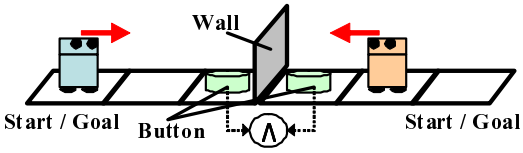


Fig. 1. Cooperative Button Pushing Problem

Table 1. Optimal local policy $\delta_i^* := (\delta_i^{A*}, \delta_i^{M*})$ for agent i on SLM in CBP problem, when $M_i = \{1, 2\}$. We assume that the initial message is $1 \in M_i$.

$\Omega_i \times \boldsymbol{M}_i^{recv}$	(SG,1)	(SG,2)	(C,1)	(C,2)	(B,1)	(B,2)
δ_i^{A*}	Fore	Fore	Fore	Back	Back	Back
δ_i^{M*}	1	1	1	2	2	2

despite for any index $i \in I$*,*

$$|M_i| < \max_{j \in I} \max_{o \in \Omega_j} |S_j^{obs}(o)|,$$

where T *and* s_0 *are the time horizon and the initial state in the Dec-POMDP-Com, respectively.*

Proof: CBP problem is a concrete example. It is a deterministic Dec-POMDP-Com with constant message cost, which is not jointly fully observable, where $I := \{1, 2\}$; for any $i \in I$, $\Omega_i := \{SG, C, B\}$, and $A_i := \{Fore, Back\}$; $S := \Omega_1 \times \Omega_2 \times \{ON, OFF\}$. Clearly, the minimum steps to reach the goal is 4. Thus, when $|M_1| = |M_2| = 2$, there exists an optimal joint policy on SLM as shown in Table 1. That is,

$$\max_{\boldsymbol{\delta} \in \boldsymbol{D}^{SLM}} V_{\boldsymbol{\delta}}^T(s_0) = \max_{\boldsymbol{\delta}' \in \boldsymbol{D}} V_{\boldsymbol{\delta}'}^T(s_0).$$

However, for any index $i \in I$,

$$\max_{o \in \Omega_i} |S_i^{obs}(o)| = 6,$$

since $S_i^{obs}(o) = \Omega_j \times \{ON, OFF\}$ for any $o \in \Omega_i$, with $j \in I : j \neq i$. ■

Fact 1 implies that an optimal joint policy on SLM can achieve the value of an optimal joint policy on the standard definition, even if it identifies the states where the optimal actions are the same with respect to an observation. Let us denote the $S_i^{obs}(o)$ shrunk by such identification by

$$A_i^{obs}(o) := \{a \in A_i \mid \delta_i^{opt}(s) = a, s \in S_i^{obs}(o)\},$$

where $\delta_i^{opt}(s)$ represents the optimal action in a global state $s \in S$. Clearly, $A_i^{obs}(o) \subseteq S_i^{obs}(o)$. Therefore, we obtain a better bound for the size of the message set M_i of each agent i as follows.

Corollary 2. *For any deterministic Dec-POMDP-Com with constant message cost, if the size $|M_i|$ of the message set of each agent i satisfies the condition,*

$$|M_i| \geq \max_{i \in I} \max_{o \in \Omega_i} |A_i^{obs}(o)|,$$

then the value of the optimal joint policy on SLM is equal to the value of any joint policy on the standard definition. That is

$$\max_{\boldsymbol{\delta} \in \boldsymbol{D}^{SLM}} V_{\boldsymbol{\delta}}^T(s_0) = \max_{\boldsymbol{\delta}' \in \boldsymbol{D}} V_{\boldsymbol{\delta}'}^T(s_0),$$

where T and s_0 are the time horizon and the initial state in the Dec-POMDP-Com.

The next fact means that SLM is truly superior to SL in a certain Dec-POMDP-Com.

Fact 2 *There exists a deterministic Dec-POMDP-Com with constant message cost, such that*

$$\max_{\boldsymbol{\delta} \in \boldsymbol{D}^{SLM}} V_{\boldsymbol{\delta}}^T(s_0) > \max_{\boldsymbol{\delta}' \in \boldsymbol{D}^{SL}} V_{\boldsymbol{\delta}'}^T(s_0),$$

where T and s_0 are the time horizon and the initial state in the Dec-POMDP-Com.

Proof: Let us consider CBP again. Since each agent on SL sends a message depending only on its own current observation, there is no more effective communication policies than sending the current observation. Thus, the agents on SL cannot obtain the status of the buttons. This means that the optimal action of each agent at the C grid is the probabilistic movement of forward and backward with equal probabilities. Therefore, an optimal policy on SL cannot achieve the value of an optimal policy on SLM. ■

5 Discussion

In Section 4, we obtained the minimum required sizes of the set of messages on SL and SLM in a decision theoretic context. In this section, we consider the minimum required sizes from the view point of reinforcement learning. Suppose that a learning algorithm converges to an optimal policy with or without the guarantee of its convergence. It is expected that we do not need the messages whose size is larger than the minimum required size, since the performance of learning may decrease as the input space of policy spreads out by increasing the extra messages. However, our experimental results show the disappointing responses to our expectation.

Here, we refer to the actual learning results of SL and SLM in CBP problem shown in Fig. 2 from our previous work [2]. We also add additional experiments shown in Fig. 3 for better understanding of the learning processes. First, we observed a strange fact from Fig. 2 that RW performed better than NC. This

fact shows that CBP problem has sufficient difficulty on the learning without communication. In other words, CBP problem is an appropriate problem for evaluating the effects of SL and SLM. By comparing the three cases except RW in Fig. 2, SL is clearly better than NC and SLM is the best in performance. In addition, SLM is more robust than SL from Fig. 3. This suggests that some beneficial meaning emerges in messages through the learning processes in SL and SLM (e.g., the agents' observations and/or the status of buttons), while in NC, the agents can receive no beneficial information as messages. The difference of performance between SL and SLM arises from the included information in messages. Although each agent on SL can only include its own observation in a message by using the local communication policy $\delta_i^M : \Omega_i \to M_i$, the local communication policy $\delta_i^M : \Omega_i \times \boldsymbol{M}_i^{recv} \to M_i$ can additionally allows each agent on SLM to include the status of buttons in a message [2]. As a result, SLM performed better than SL since the agents can deterministically decide the optimal actions based on the current global states.

Here, we go back to the main subject. In the case of SLM, the minimum required size in CBP problem is 2 from Corollary 2. The results on SLM show that the performance of when $|M_i| > 2$ is clearly better than that of when $|M_i| = 2$. This suggests that the extra messages make some positive effect in the learning. In the case of SL, we conjecture that the minimum required size in CBP problem is 3 from Corollary 1, although CBP problem is not a Dec-MDP-Com. The results on SL show that the performance of when $|M_i| > 3$ is clearly better than that of when $|M_i| = 3$. This also supports our suggestion. We currently have no credible answers about what is the positive effect. We will try it in future work.

6 Conclusions and Future Work

We defined a new model called *deterministic* Dec-POMDP-Com in Definition 10. We theoretically analyzed Signal Learning (SL) in a Dec-MDP-Com and Signal Learning with Messages (SLM) in a deterministic Dec-POMDP-Com, which we previously proposed, and obtained the minimum required sizes of the set of messages on SL and SLM in a decision theoretic context. In addition, we reviewed some experimental results, which indicates that extra messages make positive effect in learning processes, when the size of message set is larger than the minimum required size. Future work includes an extension to the stochastic model from a theoretical point of view and clarifying what is the positive effect.

References

1. Kasai, T., Tenmoto, H., Kamiya, A.: Learning of Communication Codes in Multi-Agent Reinforcement Learning Problem. In: Proceedings of the 2008 IEEE Conference on Soft Computing in Industrial Applications (SMCia/08). (2008) 1–6
2. Kasai, T., Kobayashi, H., Shinohara, A.: Improvement of the performance using Received Messages on Learning of Communication Codes. In: Proceedings of the

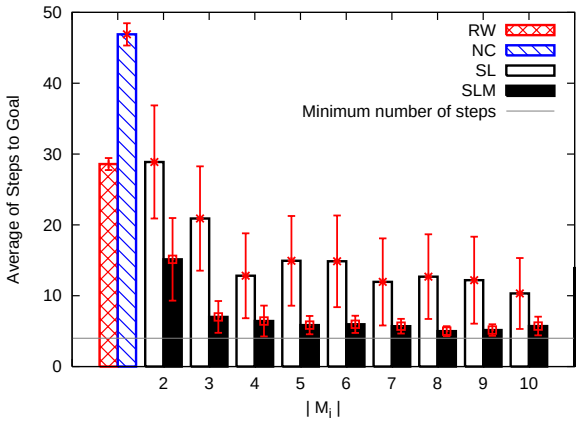
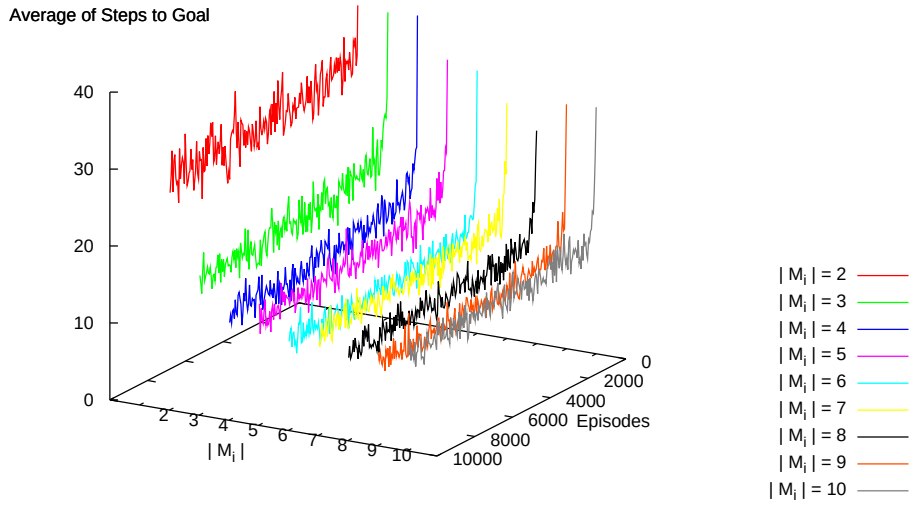
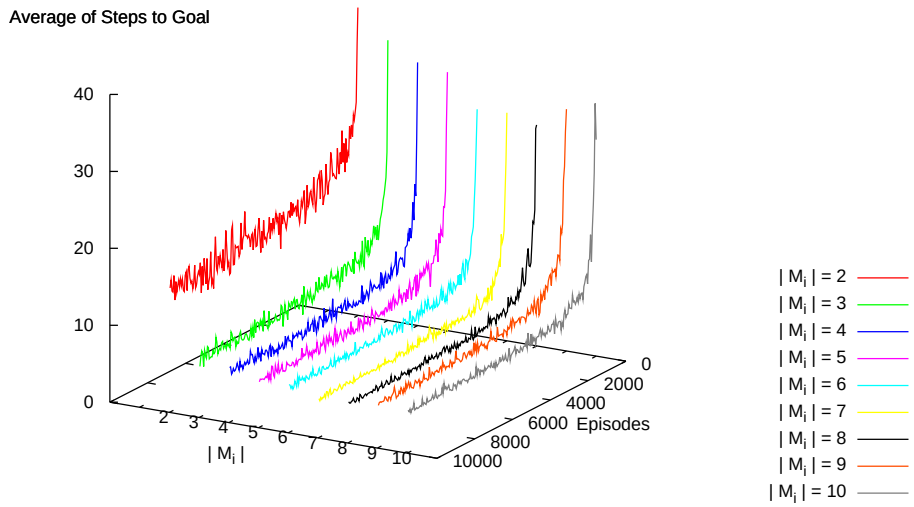


Fig. 2. This graph shows learning results averaged in 100 trials on the CBP problem, where one trial is 10,000 repetition of one episode. We used a Profit Sharing algorithm with a discount rate $\gamma = 0.5$, a reward at the goal $r = 100$. The size $|M_i|$ of the message set is varied from 2 to 10. The averaged value is the number of steps to reach the goal in the last 100 episodes in 10,000 episodes in one trial. The error bar represents the standard deviation of the corresponding bar. NC and RW represents learning with No Communication and Random Walk respectively.

8th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009), IFAAMAS (2009) 1229–1230
3. Tan, M.: Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents. In: Proceedings of the 10th International Conference on Machine Learning, Morgan Kaufmann (1993) 330–337
4. Ghavamzadeh, M., Mahadevan, S.: Learning to Communicate and Act using Hierarchical Reinforcement Learning. In: Proceedings of the 3rd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2004), IEEE Computer Society (2004) 1114–1121
5. Roth, M., Simmons, R., Veloso, M.: Decentralized Communication Strategies for Coordinated Multi-Agent Policies. In: Multi-Robot Systems: From Swarms to Intelligent Automata, volume IV, Kluwer Avademic Publishers (2005)
6. Roth, M., Simmons, R., Veloso, M.: Reasoning About Joint Beliefs for Execution–Time Communication Decisions. In: Proceedings of the 4th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2005), ACM Press (2005) 786–793
7. Roth, M., Simmons, R., Veloso, M.: What to Communicate? Execution-time Decision in Multi-agent POMDPs. In: Proceedings of the 8th International Symposium on Distributed Autonomous Robotic Systems (DARS). (2006)
8. Zhang, C., Abdallah, S., Lesser, V.: Efficient Multi-Agent Reinforcement Learning through Automated Supervision. In: Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2008), IFAAMAS (2008) 1365–1370
9. Ponsen, M.J.V., Croonenborghs, T., Tuyls, K., Ramon, J., Driessens, K.: Learning with whom to communicate using relational reinforcement learning. In: Pro-



(a) SL



(b) SLM

Fig. 3. The graph shows the variability of average steps to goal through the learning process in Fig. 2, where the averaged value in each episode is the average number of steps to reach the goal in 100 trials. The size $|M_i|$ of the message set is varied from 2 to 10.

ceedings of the 8th International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2009), IFAAMAS (2009) 1221–1222
10. Jim, K.C., Giles, C.L.: How Communication Can Improve the Performance of Multi-Agent Systems. In: Proceedings of the 5th International Conference on Autonomous Agents, ACM Press (2001) 584–591
11. Giles, C.L., Jim, K.C.: Learning Communication for Multi-agent Systems. In: Proceedings of the 1st International Workshop on Radical Agent Concepts. Volume 2564 of LNAI., Springer-Verlag (2003) 377–390
12. Goldman, C.V., Zilberstein, S.: Optimizing Information Exchange in Cooperative Multi-agent Systems. In: Proceedings of 2nd International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS 2003), ACM Press (2003) 137–144
13. Goldman, C.V., Zilberstein, S.: Decentralized Control of Cooperative Systems: Categorization and Complexity Analysis. Journal of Artificial Intelligence Research **22** (2004) 143–174
14. Allen, M., Goldman, C.V., Zilberstein, S.: Learning to Communicate in Decentralized Systems. In: Proceedings of the Workshop on Multiagent Learning at AAAI-05. (2005) 1–8
15. Goldman, C.V., Allen, M., Zilberstein, S.: Learning to communicate in a decentralized environment. Autonomous Agents and Multi-Agent Systems **15**(1) (2007) 47–90
16. Oliehoek, F.A., Spaan, M.T.J., Vlassis, N.: Dec-POMDPs with delayed communication. In: Proceedings of the AAMAS 2007 Workshop on Multi-Agent Sequential Decision Making in Uncertain Domains. (2007)
17. Seuken, S., Zilberstein, S.: Formal models and algorithms for decentralized decision making under uncertainty. Autonomous Agents and Multi-Agent Systems **17**(2) (2008) 190–250
18. Pynadath, D.V., Tambe, M.: The Communicative Multiagent Team Decision Problem: Analyzing Teamwork Theories and Models. Journal of Artificial Intelligence Research **16** (2002) 389–423