State Aggregation for Multiagent Communication over Rate-Limited Channels

Arsham Mostaani, Thang X. Vu, Symeon Chatzinotas, Björn Ottersten
Emails: arsham.mostaani, thang.vu, symeon.chatzinotas, bjorn.ottersten@uni.lu
Centre for Security Reliability and Trust, University of Luxembourg, Luxembourg

***Abstract*—A collaborative task is assigned to a multiagent system (MAS) in which agents are allowed to communicate. The MAS runs over an underlying Markov decision process and its task is to maximize the averaged sum of discounted one-stage rewards. Although knowing the global state of the environment is necessary for the optimal action selection of the MAS, agents are limited to individual observations. The inter-agent communication can tackle the issue of local observability, however, the limited rate of the inter-agent communication prevents the agent from acquiring the precise global state information. To overcome this challenge, agents need to communicate their observations in a compact way such that the MAS compromises the minimum possible sum of rewards. We show that this problem is equivalent to a form of rate-distortion problem which we call the task-based information compression. State Aggregation for Information Compression (SAIC) is introduced here to perform the task-based information compression. The SAIC is shown, conditionally, to be capable of achieving the optimal performance in terms of the attained sum of discounted rewards. The proposed algorithm is applied to a rendezvous problem and its performance is compared with two benchmarks; (i) conventional source coding algorithms and the (ii) centralized multiagent control using reinforcement learning. Numerical experiments confirm the superiority and fast convergence of the proposed SAIC.**

***Index Terms*—Task-based information compression, machine learning for communication, multiagent systems, reinforcement learning.**

I. Introduction

This paper considers a collaborative task problem composed of multiple agents with local observations, while agents are allowed to communicate through a rate limited channel. The global state process of the environment, generated by a Markov decision process (MDP), is controlled by the joint actions of the agents. Moreover, the instantaneous reward signal to which all agents have access, is influenced by the global state and agents' joint actions.

On one hand, maximizing the finite-horizon sum of discounted rewards, considered to be the unique goal of the network of agents, spurs them to act collaboratively. On the other hand, limited observability of the environment encourages the agents to effectively communicate to each other to acquire a better estimate of the global state of the environment. Due to the limited rate of the communication channel between the agents, it is necessary for agents to compactly represent their observations in communication messages in such a way that it incurs minimal compromise on the cumulative rewards. As such, this form of information compression which we call task-based information compression is different from conventional compression algorithms whose ultimate aim is to reduce the distortion between the original and compressed data.

One potential area of application for the considered framework can be object tracking by Unmanned Area Vehicle networks or by multi-agent systems, e.g. [1], [2], in which multiple UAVs/agents collaboratively track one/several moving object(s), where inter-agent communication has a rate budget. Another application for our problem is the rendezvous problem, drawn from computer science community [3], [4], where multiple agents, e.g. autonomous robots, want to get into a particular location at precisely the same time. The agents are unaware of the initial locations of each other but are allowed to communicate through a rate limited communication channel. The team of agents is rewarded if they achieve the task of arrival to the goal point at the same time, and will be punished if any of them arrives earlier.

The given examples fall in the general category of multi-agent reinforcement learning, which is used in the literature as an effective framework to develop coordinated policies [5]–[9]. The distributed decision-making of multi-agent systems has been addressed in [5], [6], while many other works are focused on multi-agent (MA) communications to enhance the joint action selection in partially observed environments [7]–[10]. Here we elaborate on some papers with focus on multi-agent communication. The work done in [7] has addressed the coordination of multiple agents through a noise-free communication channel, where the agents follow an engineered communication strategy. Deep reinforcement learning with communication of the gradients of the agents' objective function is proposed in [8] to learn the communication among multiple agents. In contrast to the above mentioned works, the impact of channel noise in the inter-agent communications is studied by [9] and the absence of dedicated communication channels by [11].

In this work, we develop a state aggregation algorithm which enables each agent to reduce the entropy of its generated communication messages while maintaining their performance in the collaborative task. Classical state aggregation algorithms have been often used to reduce the complexity of the dynamic

programming problems over MDPs [12], [13]. To the best of our knowledge, they have never been used to design a task-based information compression algorithm over an MDP. In our problem, agents' observations stem from a generative process with memory, an MDP. In contrast to the conventional design of the communication systems, we demonstrate the potential of considering the joint design of the source coding (compression) together with the multi-agent action policy design. Our particular approach is based on an indirect data-driven design exploiting multi-agent reinforcement learning.

The contributions of this paper are as follows. Firstly, we develop a general cooperative MA framework in which agents interact over an MDP environment. Unlike the existing works which assume perfect communication links [3], [8], we assume the practical rate-limited communications between the agents. Secondly, we decouple the decentralized cooperative multi-agent problem to two decentralized problems of action policy selection and communication policy selection. After transforming the communications policy selection problem into a so-called task-based rate distortion problem, we propose a state aggregation information compression as the solution. SAIC leverages centralized learning to find optimal communication policy and converts the task-based rate distortion problem to a K-median problem. Finally, the performance of the SAIC and two benchmark schemes, namely centralized control of the MAS and conventional information compression, are compared in a rendezvous problem. Note that bold font will be used for random variables and their realizations follows standard font.

II. System Model

We consider a two-agent system, where at any time step t each agent $i \in \{1, 2\}$ makes a local observation $\mathsf{o}_i(t) \in \Omega$ on environment while the true state of environment is $\mathsf{s}(t) \in \mathcal{S}$. The alphabets Ω and $\mathcal{S}$ define observation space and state space, respectively. The true state of the environment $\mathsf{s}(t)$ is controlled by the joint actions $\mathsf{m}_i(t), \mathsf{m}_j(t) \in \mathcal{M}$ of the agents, where each agent i can only choose its local action $\mathsf{m}_i(t)$ which is selected from the local action space $\mathcal{M}$. The environment runs on discrete time steps $t = 1, 2, ..., M$, where at each time step, each agent i selects its action $\mathsf{m}_i(t)$ upon having an observation $\mathsf{o}_i(t)$ of environment. State transition of the environment, conditioned on the joint actions of the MAS, are captured by a conditional probability mass function $T\big(\mathsf{s}(t+1), \mathsf{s}(t), \mathsf{m}_i(t), \mathsf{m}_j(t)\big)$, which is unknown to the agents. We recall that domain level actions $\mathsf{m}_i(t)$ can, for instance, be in the form of a movement or acceleration in a particular direction or any other type of action depending on the domain of the cooperative task. We consider a particular structure for agents' observations, referred to as collective observations in the literature [10]. Namely, at all time steps t agents' observation processes $\mathbf{o}_i(t), \mathbf{o}_j(t)$ follow eq. (1) and eq. (2).

$$H\big(\mathbf{o}_i(t)\big) \leq H\big(\mathbf{s}(t)\big), \quad i \in \{1, 2\}, \tag{1}$$

$$H\big(\mathbf{o}_i(t), \mathbf{o}_j(t)\big) = H\big(\mathbf{s}(t)\big), \quad j \neq i. \tag{2}$$

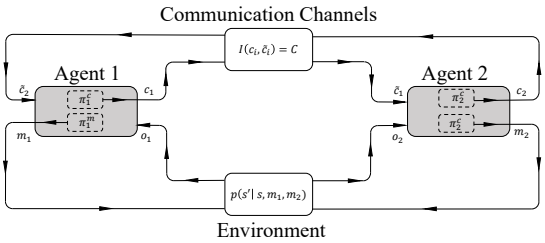

Figure 1. An illustration of the decentralized cooperative MAS with rate-limited inter-agent communications.

A deterministic reward function $r(\cdot) : \mathcal{S} \times \mathcal{M}^2 \to \mathbb{R}$ indicates the reward of both agents at time step t, where the arguments of the reward function are the global state of the environment $\mathsf{s}(t)$ and the domain-level actions $\mathsf{m}_i(t), \mathsf{m}_j(t)$ of both agents. We assume that the environment in which agents interact can be defined in terms of an MDP determined by the tuple $\big\{\mathcal{S}, \mathcal{M}^2, r(\cdot), \gamma, T(\cdot)\big\}$, where the scalar $\gamma \in [0, 1]$ is the discount factor. The focus of this paper is on scenarios in which the agents are unaware of the state transition probability function $T(\cdot)$ and of the closed form of the function $r(\cdot)$. However we assume that, further to the literature of reinforcement learning [14], a realization of the function $r\big(\mathsf{s}(t), \mathsf{m}_i(t), \mathsf{m}_j(t)\big)$ will be accessible for both agents i and j at each time step t.

In what follows the decentralized problem of MA communications and control is detailed. The main intention of this paper is to address the decentralized control and inter-agent communications for a system of multiple agents, Fig. 1. However, we keep comparing the results obtained by solving the decentralized problem with those of the centralized problem. These problems are the same in essence with the caveat that in the centralized problem perfect communications is assumed to be in place and joint actions are selected by the central controller. The optimal solution to this problem $\pi^*(\cdot) : \mathcal{S} \times \mathcal{M}^2 \to [0, 1]$ can be obtained by Q-learning [14]. Therefore, the cumulative discounted reward performance of $\pi^*(\cdot)$ can be considered as an upper-bound for the performance of the decentralized policies.

A. Problem Statement

Here we consider a scenario in which the same objective function explained in Eq. (5) needs to be maximized by the two-agent system in a decentralized fashion, Fig. 1. Namely, agents with partial observability can only select their own actions. To prevail over the limitations imposed by the local observability, agents are allowed to have direct (explicit) communications. However, the communication is done through a channel with limited rate $C = I\big(\mathbf{c}_i(t), \tilde{\mathbf{c}}_i(t)\big)$, where $\mathbf{c}_i(t) \in \mathcal{C} = \{-1, 1\}^B$ stands for the communication message generated by agent i before being encoded by error correction codes and $\tilde{\mathbf{c}}_i(t) \in \{-1, 1\}^B$ corresponds to the same communication message after it is decoded by the decoder of the same error correction code used at the transmitter, at the agent j. Note that instead of the maximal achievable rate of the channel, C represents the channel maximal (non-)asymptotic achievable rate with any known channel coding

scheme (of arbitrary code length). It should be noted that the design of the channel coding and modulation schemes are beyond the scope of this paper and the main focus is on the compression of agents' generated communication messages. In this problem, the limited rate of the channel is accepted as a constraint which is imposed by the given channel coding, length of code-words, modulation scheme and the available bandwidth. Here we assume, the achievable rate of information exchange for both inter-agent communication channels to be equal to C, i.e., the communication resources are split evenly amongst the two agents. In particular we consider C to be time-invariant and to follow:

$$\begin{cases} C < H\big(\mathbf{o}_i(t)\big), \\ C < H\big(\mathbf{o}_j(t)\big). \end{cases} \tag{3}$$

To have a more compact notation to refer to the system trajectory, hereafter, we represent the realization of a system trajectory at time t by $\mathsf{tr}(t)$ which corresponds to the tuple $\langle \mathsf{o}_i(t), \mathsf{o}_j(t), \mathsf{m}_i(t), \mathsf{m}_j(t) \rangle$ and the realization of the whole system trajectory by $\{\mathsf{tr}(t)\}_{t=1}^{t=M}$. Also for the convenience of our notation we define the function $\mathbf{g}(t^{'})$ as follows:

$$\mathbf{g}(t^{'}) = \sum_{t=t^{'}}^{M} \gamma^{t-1} r\big(\mathbf{o}_i(t), \mathbf{o}_j(t), \mathbf{m}_i(t), \mathbf{m}_j(t)\big). \tag{4}$$

Note that $\mathbf{g}(t^{'})$ is random variable and a function of $t^{'}$ as well as the trajectory $\{\mathsf{tr}(t)\}_{t=t^{'}}^{t=M}$. Due to the lack of space, here we drop a part of arguments of this function. Accordingly, the decentralized problem is formalized as

$$\begin{aligned} \max_{\pi_i^m, \pi_i^c} \quad & \mathbb{E}_{p_{\pi_i^m, \pi_i^c}\left(\{\mathsf{tr}(t)\}_{t=1}^{t=M}\right)} \Big\{ \mathbf{g}(1), \Big\} \quad i \in \{1,2\}\ , i \neq j \\ \text{s.t.} \quad & I\big(\mathbf{c}_j(t); \tilde{\mathbf{c}}_j(t)\big) \leq C, \end{aligned} \tag{5}$$

where in its general form, the action policy $\pi_i^m : \mathcal{M} \times \mathcal{C} \times \Omega \rightarrow [0,1]$ of each agent i is defined as

$$\pi_i^m\big(\mathsf{m}_i(t) | \mathsf{o}_i(t), \tilde{\mathsf{c}}_j(t)\big) = p\big(\mathsf{m}_i(t) | \mathsf{o}_i(t), \tilde{\mathsf{c}}_j(t)\big),$$

and the communication policy $\pi_i^c : \mathcal{C}^2 \times \Omega \rightarrow [0,1]$ of each agent i can be defined similarly.

As a result, the joint probability mass function of $\mathsf{tr}(1), \mathsf{tr}(2), ..., \mathsf{tr}(M)$ when each agent $i \in \{1,2\}$ follows the action policy $\pi_i^m(\cdot)$ and the communication policy $\pi_i^c(\cdot)$ is shown as $p_{\pi_i^m, \pi_i^c}\big(\{\mathsf{tr}(t)\}_{t=1}^{t=M}\big)$. The initial state $\mathbf{s}(1) \in \mathcal{S}$ is randomly selected by the environment. To make the problem more concrete, here we assume the presence of an instantaneous communication between agents [9]. Fig. 2 demonstrates this communication model. As such, each agent i at any time step t prior to the selection of its action $\mathsf{m}_i(t)$ receives a communication message $\tilde{\mathsf{c}}_j(t)$ that includes some information about the observations of agent j at time t. Under the instantaneous communication scenario, the generation of the communication message $\mathsf{c}_i(t)$ by an agent i cannot be conditioned on the received communication message $\tilde{\mathsf{c}}_j(t)$ from agent j, as it causes an infinite regress. Moreover, for agent j there will be no new information in $\tilde{\mathsf{c}}_i(t) \sim \pi_i^c\Big(\mathsf{c}_i(t) \Big| \mathsf{o}_i(t), \tilde{\mathsf{c}}_j(t)\Big)$ given $\tilde{\mathsf{c}}_i(t) \sim \pi_i^c\Big(\mathsf{c}_i(t) \Big| \mathsf{o}_i(t)\Big)$, as agent j has already full access to its own observation $\mathsf{o}_j(t)$.

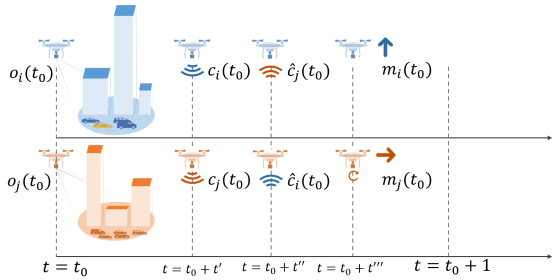

Figure 2. Illustration of instantaneous communication for a UAV object tracking example, with $0 < t' < t'' < t''' < 1$. At time $t = t_0$ agents (UAVs) make local observations. At time $t = t_0 + t'$ both agents select a communication signal. At time $t = t_0 + t''$ agents receive a communication signal. At time $t = t_0 + t'''$ agents select an domain level action.

III. State Aggregation for Information Compression in Multiagent Coordination Tasks

This section tackles the constraint on the rate of inter-agent information exchange in the problem (5) by introducing state aggregation for compression of agents observations. State aggregation in this paper is applied as a method to carry out a task-based information compression. We design the state aggregation algorithm such that it can suppress a part of the observation information that results in the smallest possible loss in the performance of the multi-agent system, where this loss is measured in terms of regret from maximum achievable expected cumulative rewards. Similar to some other recent papers with focus on multi-agent coordination [6], [8], here a centralized training phase for the two-agent system is required, however, the execution can be done in a decentralized fashion.

Here we assume that the communication resources are split evenly amongst the two agents, by considering the achievable rate of information exchange of both communication channels to be equal to C. As such, both agents compress their observations to acquire communication messages of equal entropy. For the current work also assume observations of both agents to have equal entropy $H\big(\mathsf{o}_i(t)\big) = H\big(\mathsf{o}_j(t)\big)$ and we postpone the study of non-symmetric observations and communications to the future works.

To solve problem (5), we first solve

$$\begin{aligned} \max_{\pi_i^m} \quad & \mathbb{E}_{p_{\pi_i^m, \pi_i^c}\left(\{\mathsf{tr}(t)\}_{t=1}^{t=M}\right)} \Big\{ \mathbf{g}(1) \Big\}, \quad i \in \{1,2\}\ , i \neq j \\ \text{s.t.} \quad & I\big(\mathbf{c}_j(t); \tilde{\mathbf{c}}_j(t)\big) \leq C, \end{aligned} \tag{6}$$

where we assume a general policy $\pi_i^c(\cdot)$ being followed by each agent i. Afterwards, the obtained solution for (6) can be plugged into (5) which leaves us with only one policy function $\pi_i^c(\cdot)$ to be optimized. If both of the mentioned problems can be solved optimally, then the problem (5) has been separable and the obtained solutions are the optimal solutions of it.

Accordingly, the objective function of the decentralized problem (5) can also be written as

$$\mathbb{E}_{p_{\pi_i^m,\pi_i^c}(\{\mathsf{tr}(t)\}_{t=1}^{t=M})}\{\mathbf{g}(1)\} = \mathbb{E}_{p_{\pi_i^m,\pi_i^c}(\mathsf{o}_i(1),\tilde{\mathsf{c}}_j(1))}\Big\{V_{\pi_i^m,\pi_i^c}\big(\mathbf{o}_i(1),\tilde{\mathbf{c}}_j(1)\big)\Big\}, \quad (7)$$

where $V_{\pi_i^m,\pi_i^c}\big(\mathsf{o}_i(t),\tilde{\mathsf{c}}_j(t)\big)$ is the unique solution to the Bellman equation corresponding to the joint action and communication policies π_i^m, π_i^c of both agents.

In light of eq. (7) the objective function of the problem (6) can be expressed as

$$\mathbb{E}_{p_{\pi_i^m,\pi_i^c}\left(\mathsf{o}_i(1),\tilde{\mathsf{c}}_j(1)\right)}\Big\{V_{\pi_i^m,\pi_i^c}\big(\mathbf{o}_i(1),\tilde{\mathbf{c}}_j(1)\big)\Big\}, \ i \in \{1,2\}, i \neq j. \quad (8)$$

Lemma 1, lets us to obtain the solution of (8) by finding the optimal value function $V_{\pi_i^{m*},\pi_i^c}\big(\mathbf{o}_i(1),\tilde{\mathbf{c}}_j(1)\big)$. This function can be found either by applying Bellman optimality equations for a sufficient number of times on $V_{\pi_i^m,\pi_i^c}\big(\mathbf{o}_i(1),\tilde{\mathbf{c}}_j(1)\big)$ or by Q-learning. It is important, however, to note that the value function $V_{\pi_i^{m*},\pi_i^c}\big(\mathbf{o}_i(1),\tilde{\mathbf{c}}_j(1)\big)$ obtained by Q-learning will be optimal only if, the environment that can be explained by the tuple $\Big\{\Omega \times \mathcal{C}, \mathcal{M}^2, r(\cdot), \gamma, T'(\cdot)\Big\}$ can be proven to be an MDP. Accordingly, we assume that the aggregated MDP denoted by $\Big\{\Omega \times \mathcal{C}, \mathcal{M}^2, r(\cdot), \gamma, T'(\cdot)\Big\}$ which is obtained by doing state aggregation on the original MDP denoted by $\Big\{\Omega^2, \mathcal{M}^2, r(\cdot), \gamma, T(\cdot)\Big\}$ is an MDP itself. The proof of theorems and lemmas are skipped due to the space limitation and will be available in the extended version.

Lemma 1. *The maximum of expectation of value function $V_{\pi_i^m,\pi_i^c}\big(\mathbf{o}_i(t),\tilde{\mathsf{c}}_j(t)\big)$, over the joint distribution of $\mathsf{o}_i(t),\tilde{\mathsf{c}}_j(t)$ is equal to the expectation of value function of optimal policy*

$$\max_{\pi_i^m} \quad \mathbb{E}_{p_{\pi_i^m,\pi_i^c}\left(\mathsf{o}_i(1),\tilde{\mathsf{c}}_j(1)\right)}\Big\{V_{\pi_i^m,\pi_i^c}\big(\mathbf{o}_i(1),\tilde{\mathbf{c}}_j(1)\big)\Big\} = \mathbb{E}_{p_{\pi_i^m,\pi_i^c}\left(\mathsf{o}_i(1),\tilde{\mathsf{c}}_j(1)\right)}\Big\{V_{\pi_i^{m*},\pi_i^c}\big(\mathbf{o}_i(1),\tilde{\mathbf{c}}_j(1)\big)\Big\} \quad (9)$$

Remember that numerical methods such as value iteration or Q-learning, cannot normally provide parametric solutions which is in contrast to our requirements in SAIC, as explained earlier in this section. Lemma 2, allows us to acquire a parametric approximation of $V_{\pi_i^{m*},\pi_i^c}\big(\mathbf{o}_i(1),\tilde{\mathbf{c}}_j(1)\big)$ by leveraging the value function $V^*\big(\mathsf{o}_i(t),\mathsf{o}_j(t)\big)$ corresponding to the optimal solution of the centralized problem. Accordingly, following lemma 2, we propose to derive an off-policy approximation of $V_{\pi_i^{m*},\pi_i^c}\big(\mathbf{o}_i(t),\tilde{\mathbf{c}}_j(t)\big)$ having the knowledge of policy $\pi^*(\cdot)$.

Lemma 2. *The optimal value of $V^*\big(\mathsf{o}_i(t),\tilde{\mathsf{c}}_j(t)\big)$ can be obtained using the solution $\pi^*(\cdot)$ and its corresponding value function $V^*\big(\mathsf{o}_i(t),\mathsf{o}_j(t)\big)$ following*

$$V^*\big(\mathsf{o}_i(t),\tilde{\mathsf{c}}_j(t)\big) = \sum_{\mathsf{o}_j(t)\in Q_k} V^*\big(\mathsf{o}_i(t),\mathsf{o}_j(t)\big)p\big(\mathsf{o}_j(t)|\tilde{\mathsf{c}}_j(t)\big). \quad (10)$$

Based on the results of lemma 1 and lemma 2, theorem 3 is constructed such that it allows us to compute the communication policies of agents independent from their action policies. The proposed communication policy by theorem 3, is conditionally the optimal communication policy.

Theorem 3. *The communication policy that can maximize the achievable expected cumulative rewards in the decentralized coordination problem (5) can be obtained by solving the k-median clustering problem*

$$\min_{\mathcal{P}_i} \quad \sum_{k=1}^{2^B} \sum_{\mathsf{o}_i(t)\in\Omega} \Big|V^*\big(\mathsf{o}_i(t)\big) - \mu_k^{'}\Big|, \quad (11)$$

where $\mathcal{P}_i$ corresponds to a unique $\pi_i^c(\cdot)$.

Theorem 3 allows us to compute a communication policy $\pi_i^c(\cdot)$ by clustering values of $V^*\big(\mathsf{o}_i(t)\big)$, where this policy can be the optimal communication policy under some conditions which are further discussed later on in this section. One way to obtain $V^*\big(\mathsf{o}_i(t)\big)$ is to solve the centralized problem by Q-learning. By solving this problem $Q^*(\mathsf{o}_1,\mathsf{o}_2,\mathsf{m}_1,\mathsf{m}_2)$ can be obtained. Accordingly, following Bellman optimality equation, we can compute $V^*\big(\mathsf{o}_i(t)\big)$ by

$$\max_{\mathsf{m}} \ Q^{e^*}(\mathsf{o}_1,\mathsf{m}) = V^*(\mathsf{o}_1), \quad (12)$$

where $V^*(\mathsf{o}_1)$ can be expressed as

$$V^*(\mathsf{o}_1) = \mathbb{E}_{\pi^*}\left\{\sum_{t=1}^{M}\gamma^{t-1}r\big(\mathbf{s}(t),\mathbf{m}_i(t),\mathbf{m}_j(t)\big)|\mathbf{o}_i(t)=\mathsf{o}_1\right\} \quad (13)$$

and further to the law of iterated expectations, can also be written as

$$V^*(\mathsf{o}_1) = \sum_{\mathsf{o}_2\in\Omega}\max_{\mathsf{m}_1,\mathsf{m}_2} Q^*(\mathsf{o}_1,\mathsf{o}_2,\mathsf{m}_1,\mathsf{m}_2)p(\mathbf{o}_j(t)=\mathsf{o}_2). \quad (14)$$

In SAIC, detailed in Algorithm 1, we first compute the value $V^*(\mathsf{o})$ for all $\mathsf{o} \in \Omega$. Afterwards, by solving the k-median clustering problem (11), an observation aggregation scheme indicated by $\mathcal{P}_i$ is computed. By following this aggregation scheme, the observations $\mathsf{o}_i(t) \in \Omega$ will be aggregate such that the performance of the multi-agent system in terms of the the objective function it attains is optimized. After obtaining the exact communication/compression policy $\pi_i^c(\cdot)$, an exact action policy for both agents corresponding to $\pi_i^c(\cdot)$ will be obtained by Q-learning. As such, the second training phase in which the action Q-tables $Q_i^m(\cdot)$ for $i = \{1,2\}$ are obtained as well as the execution phase of the algorithm can be done distributively.

IV. Numerical results

In this section, we evaluate our proposed schemes via numerical results for the popular rendezvous problem, in which the inter-agent communication channel is set to have a limited rate. To find the details of the rendezvous problem, please refer to [9]. The system operates in discrete time, with agents taking actions and communicating in each time step $t = 1, 2, \ldots$. We consider a variety of grid-worlds sizes, with

Algorithm 1 State Aggregation for Information Compression

Input: γ, α, c
Initialize all-zero table $N_i^m\big(\mathsf{o}_i(t), \tilde{\mathsf{c}}_j(t), \mathsf{m}_i(t)\big)$, for $i = 1, 2$
  and Q-table $Q_i^m(\cdot) \leftarrow Q_i^{m,(k-1)}(\cdot)$, for $i = 1, 2$
  and all-zero Q-table $Q\big(\mathsf{o}_i(t), \mathsf{o}_j(t), \mathsf{m}_i(t), \mathsf{m}_j(t)\big)$.
Obtain $\pi^*(\cdot)$ and $Q^*(\cdot)$ by solving the centralized problem [14].
Compute $V^*\big(\mathsf{o}_i(t)\big)$ following eq. (14), for $\forall \mathsf{o}_i(t) \in \Omega$.
Solve problem (11) by applying k-median clustering to obtain $\pi_i^c(\cdot)$, for $i = 1, 2$.
for each episode $k = 1 : K$ **do**
  Randomly initialize local observation $\mathsf{o}_i(t = 1)$, for $i = 1, 2$
  for $t_k = 1 : M$ **do**
    Select $\mathsf{c}_i(t)$ following $\pi_i^c(\cdot)$, for $i = 1, 2$
    Obtain message $\tilde{\mathsf{c}}_j(t)$, for $i = 1, 2 \;\; j \neq i$
    Update $Q_i^m\big(\mathsf{o}_i(t-1), \tilde{\mathsf{c}}_j(t-1), \mathsf{m}_i(t-1)\big)$, for $i = 1, 2$
    Select $\mathsf{m}_i(t) \in \mathcal{M}$ following UCB policy [14], for $i = 1, 2$
    Increment $N_i^m\big(\mathsf{o}_i(t), \tilde{\mathsf{c}}_j(t), \mathsf{m}_i(t)\big)$, for $i = 1, 2$
    Obtain reward $r\big(\mathsf{o}_i(t), \mathsf{o}_j(t), \mathsf{m}_i(t), \mathsf{m}_j(t)\big)$, for $i = 1, 2$
    Make a local observation $\mathsf{o}_i(t)$, for $i = 1, 2$
    $t_k = t_k + 1$
  end
  Compute $\sum_{t=1}^{M} \gamma^t r_t$ for the lth episode
end
Output: $Q_i^m(\cdot)$ and $\pi_i^m\big(\mathsf{m}_i(t) | \mathsf{o}_i(t), \tilde{\mathsf{c}}_j(t)\big)$
  and by following greedy UCB policy, for $i = 1, 2$

different values for N for instance $N = 4$ means a 4×4 grid-world, and different locations for the goal-point ω^T. We compare the proposed SAIC with (i) Centralized Q-learning scheme and (ii) the Conventional Information Compression (CIC) scheme. In CIC we first train the disjoint action policies using distributed Q-learning, assuming presence of ideal inter-agent communications. Subsequently, leveraging an estimated distribution of agent i's observations $\mathbf{o}_i(t)$, the observations will be quantized using Lloyd's algorithm.

A. Results

To perform our numerical experiments, rewards of the rendezvous problem are selected as $R_1 = 1$ and $R_2 = 10$, while the discount factor is $\gamma = 0.9$. A constant learning rate $\alpha = 0.07$ is applied, and the UCB exploration rate $c = 12.5$. In any figure that the performance of each scheme is reported in terms of the averaged discounted cumulative rewards, the attained rewards throughout training iterations are smoothed using a moving average filter of memory equal to 20,000 iterations. Regardless of the grid-world's size and goal location, the grids are numbered row-wise starting from the left-bottom.

Fig. 3 illustrates the performance of the proposed scheme SAIC as well as the two benchmark schemes centralized Q-learning and CIC. The grid-world is considered to be of size $N = 8$ and its goal location to be $\omega^T = 22$. The rate budget of the channel is $C = 2$ bits per time step. Since the centralized Q-learning runs with perfect communications, it achieves optimal performance after enough training, 160k iterations. The CIC, due to insufficient rate of the communication channel never achieves the optimal solution. It is observed that the SAIC by less than 1% gap achieves optimal performance and does that remarkably fast. The yellow curve showing the performance of the CIC with no communication between agents, would show us the best performance that can be achieved if no communication between agents is in place. Note that both the CIC and SAIC require a separate training phase which is not captured by Fig. 3. SAIC requires a centralized training phase and CIC a distributed training phase with unlimited capacity of inter-agent communication channels. The performance of these two algorithms in Fig. 3 is plotted after the first phase of training. To understand the underlying reasons for the

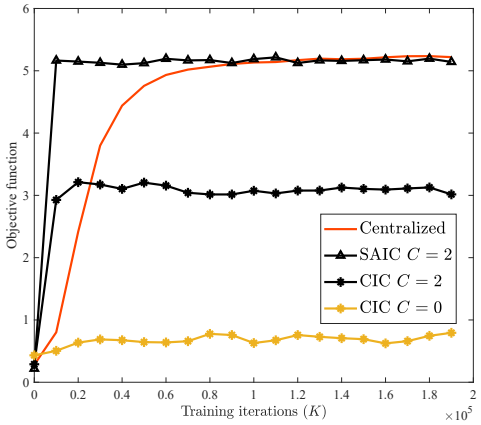

Figure 3. A comparison between all four schemes in terms of the achievable objective function with channel rate constraint $C = 2$ bits per time steps and number of training iterations/episodes $K = 200k$.

remarkable performance of the SAIC, Fig. 4 is provided so that the equivalence classes computed by the SAIC can be seen, with all the locations of the grid shaded with the same colour belonging to the same equivalence class. The SAIC is extremely efficient, in performing state aggregation such that the loss of observation information does not incur any loss of achievable sum of discounted rewards. Fig. 4-(a), illustrates the state aggregation obtained by the SAIC, for which the achievable sum of discounted rewards is illustrated in Fig. 3. It is illustrated in Fig. 4-(a) that how the SAIC performs observation compression with ratio $R = 3 : 1$, while it leads to nearly no performance loss for the collaborative task of the MAS. Here the definition of compression ratio follows $R = \lceil H\big(\mathbf{o}_i(t)\big)\rceil / \lceil H\big(\mathbf{c}_i(t)\big)\rceil$.

We also investigate the impact of achievable bit rate C on the achievable value of objective function for the SAIC and CIC, in Fig. 5. In this figure, the normalized value of achieved objective function for any scheme at any given C is shown. The average of the attained objective function for the scheme of interest is computed by $\mathbb{E}_{p_{\pi_i^m, \pi_i^c}(\{\mathsf{tr}(t)\}_{t=1}^{t=M})}\big\{\mathbf{g}(1)\big\}$, where $\pi_i^m(\cdot)$ and $\pi_i^c(\cdot)$ are obtained by the scheme of interest after solving (5). The attained objective function for the scheme of interest is then normalized by dividing it to the average of objective function $\mathbb{E}_{p_{\pi^*}(\{\mathsf{tr}(t)\}_{t=1}^{t=M})}\big\{\mathbf{g}(1)\big\}$ that is attained by

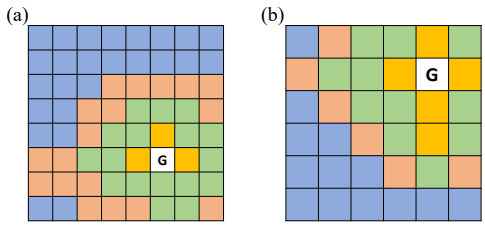

Figure 4. State aggregation for multi-agent communication in a two-agent rendezvous problem with grid-worlds of varied sizes and goal locations. The observation space is aggregated to four equivalence classes, $C = 2$ bits, and number of training episodes has been $K = 1500k$ and $K = 1000k$ for figure (a) and (b) respectively. Locations with similar color represent all the agents' observations which are grouped into the same equivalence class.

the optimal centralized policy $\pi^*(\cdot)$. Accordingly, when the normalized objective function of a particular scheme is seen to be close to the value 1, the scheme has been able to compress the observation information with almost zero loss in the achieved objective function. On one hand, it is demonstrated that the SAIC soon achieves the optimal performance, while it takes the CIC a rate of at least $C = 4$ to have near optimal performance. A whopping 40% performance gain is acquired by the SAIC, in comparison to the CIC, at high compression ratio $R = 3 : 1$, i.e., $C = 2$. This means 66% of data rate saving with no performance drop in attaining the collaborative objective function.

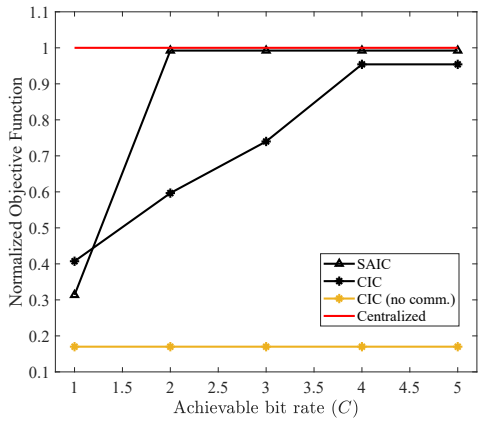

Figure 5. A comparison between the performance of several multi-agent communication and control schemes in terms of the achieved value of the objective function under different achievable bit rates. All experiments are performed on a grid-world of size $N = 8$, where the goal point is located on the grid no. 22, similar to the one depicted on Fig. 4 -a. The number of training episodes/iterations for any scheme at any given value of C has been $K = 200K$.

As was demonstrated through numerical experiments, the weakness of conventional schemes for compression of agents' observations is that they might lose/keep information regardless of how useful they are to achieve the optimal objective function. In contrast, the task-based compression scheme SAIC, for communication rates lower than the entropy of the observation process, manages to compress the observation information not to minimize the distortion but to maximize the achievable value of the objective function.

V. Conclusion

This paper has investigated a collaborative MA reinforcement learning problem. We have aimed at optimizing the MAS's objective function by means of distributed control of agents enabled with inter-agent communication. Since we consider a limited rate for the MA communication channels, task-based compression of agents observations has been of the essence. We designed SAIC with optimal performance on maximizing the achievable objective function given the constraint on the rate of communication. The proposed scheme is seen to outperform conventional source coding algorithms, by up to a remarkable 40% difference in the achieved objective function. The introduced information compression scheme, SAIC, can have a substantial impact in many communication applications, e.g. device to device communications, where the ultimate goal of communication is not a reliable transfer of information between two ends but is to acquire information which is useful to improve an achievable team objective.

References

[1] Y. Sung, A. K. Budhiraja, R. K. Williams, and P. Tokekar, "Distributed assignment with limited communication for multi-robot multi-target tracking," *Autonomous Robots*, vol. 44, no. 1, pp. 57–73, 2020.

[2] S. Li, G. Battistelli, L. Chisci, W. Yi, B. Wang, and L. Kong, "Computationally efficient multi-agent multi-object tracking with labeled random finite sets," *IEEE Trans. on Signal Processing*, vol. 67, no. 1, pp. 260–275, 2019.

[3] P. Xuan, V. Lesser, and S. Zilberstein, "Communication decisions in multi-agent cooperation: Model and experiments," in *Proc. of the Fifth Intl. Conf. on Autonomous Agents*, 2001, p. 616–623.

[4] C. Amato, J. S. Dibangoye, and S. Zilberstein, "Incremental policy generation for finite-horizon dec-pomdps," in *19th Int. Conf. on Automated Planning and Scheduling*, 2009.

[5] S. Kar, J. M. F. Moura, and H. V. Poor, "$\mathcal{QD}$-learning: A collaborative distributed strategy for multi-agent reinforcement learning through $\mathsf{m}consensus + \mathsf{m}innovations$," *IEEE Trans. on Signal Processing*, vol. 61, no. 7, pp. 1848–1862, April 2013.

[6] J. N. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *32nd AAAI Conf. on Artificial Intelligence*, 2018.

[7] C. Zhang and V. Lesser, "Coordinating multi-agent reinforcement learning with limited communication," in *Conf. on Autonomous Agents and Multi-agent Sys.*, St. Paul, Minnesota, May 2013, pp. 1101–1108.

[8] J. Foerster, Y. Assael, N. de Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *Proc. Advances in Neural Inf. Processing Systems*, Barcelona, 2016.

[9] A. Mostaani, O. Simeone, S. Chatzinotas, and B. Ottersten, "Learning-based physical layer communications for multiagent collaboration," in *2019 IEEE 30th Annual Intel. Symp. on Personal, Indoor and Mobile Radio Communications (PIMRC)*, Sep. 2019, pp. 1–6.

[10] D. V. Pynadath and M. Tambe, "The communicative multiagent team decision problem: Analyzing teamwork theories and models," *Journal of Artificial Intelligence Research*, vol. 16, pp. 389–423, Jun. 2002.

[11] B. Larrousse, S. Lasaulce, and M. R. Bloch, "Coordination in distributed networks via coded actions with application to power control," *IEEE Trans. on Information Theory*, vol. 64, no. 5, pp. 3633–3654, 2018.

[12] D. P. Bertsekas and D. A. Castanon, "Adaptive aggregation methods for infinite horizon dynamic programming," *IEEE Transactions on Automatic Control*, vol. 34, no. 6, pp. 589–598, June 1989.

[13] D. P. Bertsekas, "Feature-based aggregation and deep reinforcement learning: A survey and some new implementations," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, no. 1, pp. 1–31, 2018.

[14] R. S. Sutton and A. G. Barto, *Introduction to reinforcement learning*, 2nd ed. MIT Press, Nov. 2017, vol. 135.