Scalable Quantification of the Value of Information for Multi-Agent Communications and Control Co-design

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

***Abstract*—Task-oriented communication design (TOCD) has gained significant attention from the research community due to its numerous promising applications in domains such as IoT and industry 4.0. This paper introduces an innovative approach to designing scalable task-oriented quantization and communications in cooperative multi-agent systems (MAS). Our proposed approach leverages the TOCD framework and the concept of the value of information (VoI) to facilitate efficient communication of quantized observations among agents while maximizing the average return performance of the MAS-a metric that measures the task effectiveness of the MAS. Learning the VoI becomes a prohibitively large computational problem as the number of agents grows in the MAS. To address this challenge, we present a three-step framework. First, we employ reinforcement learning (RL) to learn the VoI for a two-agent, rather than for the original N-agent system, reducing the computational costs associated with obtaining the value of information. Next, we design the quantization policy for a MAS with N agents, utilizing the learned VoI across a range of bit-budgets. The resulting quantization strategy for agents' observations, ensures that more valuable observations are communicated with greater precision. Finally, we apply RL to learn the agents' control policies, while adhering to the quantization policies designed in the previous step. Our analytical results showcase the effectiveness of the proposed framework across a wide range of problems. Numerical experiments demonstrate improvements in reducing the computational complexity required for obtaining VoI by five orders of magnitude in TOCD for MAS problems while compromising less than 1% on the average return performance of the MAS.**

***Index Terms*—Task-oriented data compression/quantization, communications for machine learning, machine learning for communications, multiagent systems, reinforcement learning.**

I. Introduction

We are witnessing a remarkable surge in applications necessitating seamless communications between machines, representing a paradigm shift from traditional human-centred communications. While it is reasonable to design a communication pipeline aiming at minimizing the information distortion between the transmitting and receiving ends, communication between machines has a different mission which cannot be effectively captured by minimizing the traditional distortion metrics. In fact, communications for a machine at the receiving end are to improve the accuracy or timeliness of a computational task performed at the receiver. Moreover, the computational task of the receiving end has to abide by the limitations coming from the side of communications, ultimately, resulting in a joint communication and computation design. Due to the huge savings they offer in radio resource expenditure, joint design of communications and computations are receiving ever-increasing attention from the research community [1]–[7].

A certain class of the joint design of communications and computations is the joint communications and control design (JCCD) [3]–[9], where the computational task carried out at the receiver impacts the future observations of the transmitting end. While some of the existing works are focused on the JCCD problem for a single controller [6], others address the problem for a multi-agent system [3], [4], [7]. With a double exponential time complexity, however, solving the joint problem for a multi-agent system proves challenging [9]. Accordingly, some efforts are put in place to avoid directly solving the JCCD problem while designing task-aware communications [6], [7], [10] - meaning that the communication policy is designed such that it is aware of the control/computational task carried out at the receiving end. This can be done by adopting task-dependent fidelity criteria to design task-oriented data transmissions [5]. The framework proposed by this paper provides insights into how to use the value of information (VoI) as a measure to quantify the semantic/task-relevant content of an agent's observations in a MAS. Interestingly, the measure of VoI that we arrive at in this paper is similar to the one proposed by [11], with a single difference: the proposed measure in [11], quantifies the expected difference in the average return of the system with and without knowing some information, whereas by VoI, here, we refer to the expected return of the system knowing some information[1].

Motivated by the above discussions, in this paper we aim at reducing the complexity of learning VoI from exponential time complexity - $\mathcal{O}(c^N)$ with $c > 1$ - to constant time complexity - $\mathcal{O}(1)$ - with respect to the number of agents N. We show that a two-agent centralized training phase is sufficient to effectively learn the VoI function - when the reward function and observation structure follow certain conditions. Regardless of the method used in the centralized training to compute the VoI of an agent's observation e.g., deep reinforcement learning, exact reinforcement learning or dynamic programming, our analytical results stay valid. According to these

This work is supported by European Research Council (ERC) advanced grant 2022 (Grant agreement ID: 742648).

[1]We have shown before in [7], how preserving our defined VoI in communications between agents, can maintain the performance of the MAS.

results, we propose a scalable state aggregation algorithm for data compression (ESAIC) which can easily be applied to MASs composed of many agents to fulfil a collaborative task. Carrying out numerical studies on geometrical consensus problem [12], will show that the proposed ESAIC is capable of reducing the complexity of the centralized training for hundreds of days - if not years - even in very simple problems, while it maintains the average return performance of the algorithm close to the optimality.

The remainder of this paper is organized as follows. The problem of optimizing the performance of a cooperative MAS for a general task under rate-constrained inter-agent communications is cast in section II. Section III proposes the extended SAIC (ESAIC), an algorithm for the joint design of data quantization and control which enjoys reduced computational complexity and similar average return performance compared with SAIC [7]. Section IV analyzes the proposed ESAIC algorithm. Section V reports the result of the numerical experiments. Finally, the paper is concluded in Section VI.

Notations: Random variables and their realizations are differentiated via bold and simple font respectively. We also use the concept of image functions in our analytical studies which is defined as the following. Let $g(\cdot) : \mathcal{D} \to \mathcal{C}$ be a function and $\mathcal{D}' \subset \mathcal{D}$ be a subset of its domain. The image function of $g(\cdot)$ denoted by $\ddot{g}(\cdot) : \mathbb{P}(\mathcal{D}) \to \mathbb{P}(\mathcal{C})$ is defined as $\ddot{g}(\mathcal{D}') \triangleq \{c \in \mathcal{C} \mid g(d) = c \ , d \in \mathcal{D}'\}$. For the sake of the simplicity of the analysis, the arguments of the function may be omitted when no confusion is raised, e.g., we have used $r^{[n]}(\cdot)$ instead of $r^{[n]}(\mathsf{o}_1, ..., \mathsf{o}_n, \mathsf{m}_1, ..., \mathsf{m}_n)$.

II. Problem Statement

Let $\mathcal{N} = \{1, 2, ..., N\}$ be a multi-agent system composed of N agents which execute a cooperative task distributedly. The system runs on discrete time steps t. At every time step t, each agent $i \in \mathcal{N}$ observes $\mathbf{o}_i(t) \in \Omega$ while the state $\mathbf{s}(t) \in \mathcal{S}$ of the system is defined by the vector of joint observations $\mathbf{s}(t) \triangleq [\mathbf{o}_i(t)]_{i\in\mathcal{N}} \in \Omega^N$. Now let $\mathbf{s}_i(t) \in \{\Omega \cup 0\}^N$ be the vector of agent i's local state, with all its elements being equal to zero except for its i'th element which is equal to $\mathbf{o}_i(t)$. We assume that $\forall i, j \in \mathcal{N}$ the local states $\mathbf{s}_i(t)$ and $\mathbf{s}_j(t)$ are linearly independent. This is also referred to as joint observability of the state. At the time step t each agent i's control is denoted by $\mathbf{m}_i(t) \in \mathcal{M}$, and the vector of all agents' control by $\mathbf{m}(t) \in \mathcal{M}^N$ which is, in fact, a collection of all agents' controls $\mathbf{m}(t) \triangleq \langle \mathbf{m}_1(t), ..., \mathbf{m}_N(t)\rangle$. All of the observation, state and action spaces $\Omega, \mathcal{S}, \mathcal{M}$ are discrete sets. The dynamics of the environment are represented by an underlying Markov Decision Process that is denoted by the tuple $M = \langle \mathcal{S}, \mathcal{M}^N, r(\cdot), \gamma, T(\cdot)\rangle$, with $r(\cdot) : \mathcal{S} \times \mathcal{M}^N \to \mathbb{R}$ being the stage reward function, and the scalar $0 \leq \gamma \leq 1$ the discount factor. Also, the function $r^{[n]}(\cdot) : \Omega^n \to \mathcal{M}^n$ is the reward function of an MAS comprised of n agents. The function $T(\cdot) : \mathcal{S} \times \mathcal{M}^N \times \mathcal{S} \to [0, 1]$ is a conditional probability mass function (PMF) which represents state transitions such that $T\big(\mathsf{s}(t+1), \mathsf{m}(t), \mathsf{s}(t)\big) = \Pr\big(\mathsf{s}(t+1)|\mathsf{s}(t), \mathsf{m}(t)\big)$. The performance of the MAS is measured according to the system's average return defined as the summation of obtained per-stage rewards within the time horizon T':

$$\mathbf{g}(t^{'}) = \sum\nolimits_{t=t'}^{T'} \gamma^{t-1} r\big(\mathbf{s}(t), \mathbf{m}(t)\big). \tag{1}$$

Once per time step agent $i \in \mathcal{N}$ is allowed to transmit a communication vector $\mathsf{c}_i(t)$ to every other agent $j \in \mathcal{N}_{-i} = \mathcal{N}_{-i}$ - following a full mesh topology for connectivity. Conditioned on its observation $\mathsf{o}_i(t)$, agent i transmits a vector of communication messages $\mathsf{c}_i(t) = [\mathsf{c}_{i,j}(t)]_{j\in\mathcal{N}_{-i}} \in \prod_{j\in\mathcal{N}_{-i}} \mathcal{C}_{i,j}$, in which the element $\mathsf{c}_{i,j}(t)$ denotes the message sent by agent i to agent j, where $\mathsf{c}_{i,j}(t)$ is generated following the communication policy $\pi^c_{i,j}(\cdot) : \Omega \to \mathcal{C}_{i,j}$. The non-empty set $\mathcal{C}_{i,j}$ is an alphabet $\{\mathsf{c}_{i,j}, \mathsf{c}'_{i,j}, \mathsf{c}''_{i,j}, ..., \mathsf{c}^{(B_{i,j}-1)}_{i,j}\}$ composed of a finite $B_{i,j}$ number of communication code-words - we use the same notation to refer to the different elements of the action, observation and state spaces too. Agent i's communications are generated by following the tuple $\pi^c_i = \langle \pi^c_{i,j}(\cdot)\rangle_{j\in\mathcal{N}_{-i}}$ which is comprised of $N-1$ different communication policies. Agent i's communications are sent over $N-1$ separate error-free finite-rate bit pipe, with its rate constraint to be $R_{i,j} \in \mathbb{N}$ (bits per channel use) or equivalently (bits per time step). As a result, the cardinality of the communication symbol space $\mathcal{C}_{i,j}$ for each i to j inter-agent communication link should follow the inequality

$$0 \leq B_{i,j} \leq 2^{R_{i,j}}. \tag{2}$$

In the special case of homogeneous bit-budgets, we have $R_{i,j} = R, \forall i, j \in \mathcal{N}$. Each agent i exploits a combination of its local observation $\mathbf{o}_i(t)$ as well as all the received quantized messages $\tilde{\mathsf{c}}_i(t) = [\mathsf{c}_{j,i}(t)]_{j\in\mathcal{N}_{-i}} \in \prod_{j\in\mathcal{N}_{-i}} \mathcal{C}_{j,i}$ within time-step t to select the control signal $\mathsf{m}_i(t)$ following a deterministic control policy $\pi^m_i(\cdot) : \prod_{j\in\mathcal{N}_{-i}} \mathcal{C}_{j,i} \times \Omega \to \mathcal{M}$. Accordingly, the problem we solve is detailed in Definition 1.

Definition 1. ***(Distributed Joint Control and Communication Design (D-JCCD) problem).*** *Let M be the MDP governing the environment and the scalar $R_{i,j} \in \mathbb{R}$ to be the bit-budget of each inter-agent communication channels. At any time step t', we aim at designing the tuple $\pi_i = \langle \pi^m_i(\cdot), \pi^c_i\rangle$ to solve the following variational dynamic programming*

$$\underset{\pi_i}{\text{argmax}} \ \ \mathbb{E}_{\pi_i}\Big\{\mathbf{g}(t')\Big\}; \ \ \text{s.t.} \ \ B_{i,j} \leq 2^{R_{i,j}}, \ \forall i, j \in \mathcal{N} \tag{3}$$

where the expectation is taken over the joint pmf of system's trajectory $\{\mathsf{tr}\}^{T'}_{t'} = [\mathsf{o}_1(t'), ..., \mathsf{o}_N(t'), \mathsf{m}(t'), ..., \mathsf{o}_1(T'), ..., \mathsf{o}_N(T'), \mathsf{m}(T')]$, when each agent i follows the policy π_i for all agents $i \in \mathcal{N}$.

In contrast to [7], we do not characterize the performance gap caused by the limited connectivity in the communication network of agents. Characterizing the difference between the performance of the MAS that runs over heterogeneous bit-budgets and the MAS that runs over perfect communication channels is deferred to future works. The present paper, however, will provide numerical studies on the performance of the proposed scheme - ESAIC - under asymmetrical communication bit-budgets $R_{i,j}$.

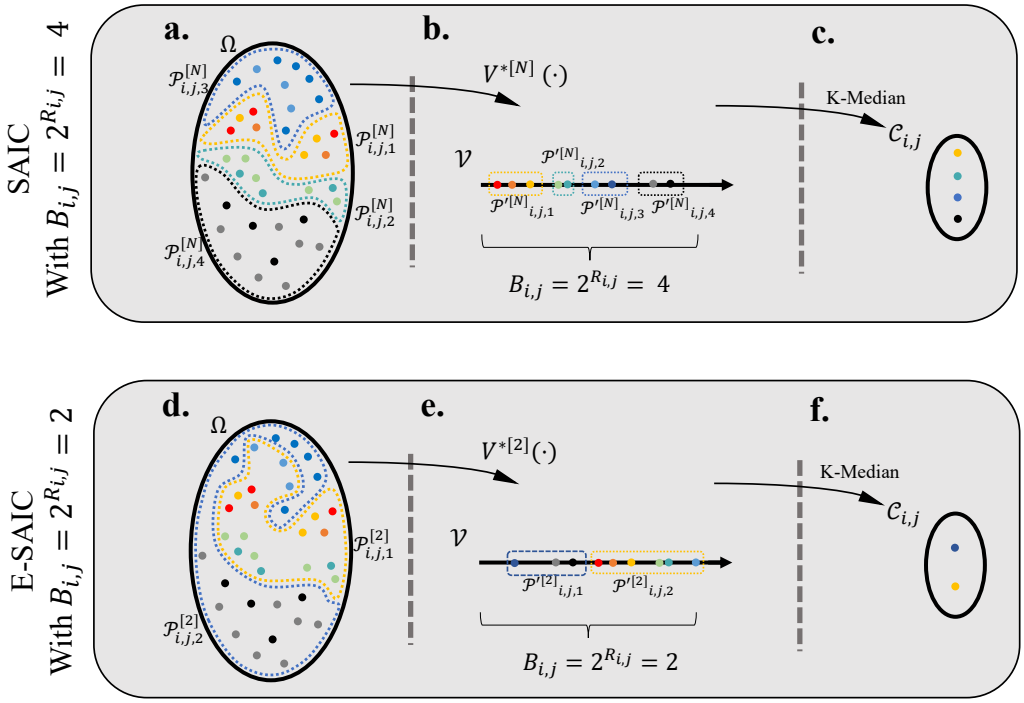

Figure 1. Illustration of the steps taken to design the communication policy $\pi^{c}_{i,j}(\cdot)$ using SAIC and ESAIC.

III. Extended State Aggregation for Information Compression in Multiagent Coordination Tasks

In this section, we propose a straightforward extension of SAIC [7], called Extended SAIC (ESAIC) which is capable of drastically reducing its time complexity in the centralized training phase to learn the VoI. While the time complexity of the centralized training phase in SAIC grows exponentially with respect to the number of agents, in ESAIC, it stays constant with respect to the number of agents N - making ESAIC more efficient than SAIC [7] and any other MARL with a central training phase [10], [13]. ESAIC is not just a replacement for SAIC, but introduces the more general idea of reducing the number of agents in the MAS for the centralized training phase. Extended SAIC, proceeds by following the same steps as SAIC to solve the D-JCCD problem: (i) centralized training phase, (ii) task-oriented data compression problem, (iii) distributed training of agents' control policies. The only difference is that the centralized training phase is done with only two agents in the training phase - regardless of the number of agents N for which we want to solve the original D-JCCD problem (3).

A. Centralized Training Phase

We perform a centralized training phase, by solving the centralized control problem (4) for a two-agent system, with $N \geq 2$, to obtain $V^{*[2]}(\cdot)$, $\pi^{*[2]}(\cdot)$ following

$$\pi^{*[2]}(\cdot) = \underset{\pi(\cdot)}{\text{argmax}} \ \mathbb{E}_{\pi}\Big\{\mathbf{g}(t)\Big\}. \tag{4}$$

B. Task-oriented communication/quantization design (TOCD)

Afterwards, by solving the following task-oriented quantization problem

$$\min_{\mathcal{P}^{[2]}_{i,j}} \ \sum\nolimits_{k=1}^{2^{R_{i,j}}} \sum\nolimits_{\mathsf{o}\in\mathcal{P}_{i,k}} \Big| V^{*[2]}\big(\mathsf{o}_i(t)=\mathsf{o}\big) - \mu^{'}_{k} \Big|, \tag{5}$$

we obtain a new partition $\mathcal{P}^{[2]}_{i,j}$ of the observation space that leads to a different, yet effective communication/quantization policy $\pi^{c[2]}_{i,j}(\cdot)$. K-median clustering can be used to solve the above-mentioned problem (5). In this direction, to obtain the quantization policy of agent i for its communication to agent j we compute a partition $\mathcal{P}'^{[2]}_{i,j}$ of the set $\mathcal{V}^{[2]}_{i}$ - where $\mathcal{V}^{[2]}_{i}$ is the image of Ω under the function $V^{*[2]}(\cdot)$ i.e., $\mathcal{V}^{[2]}_{i} = \ddot{V}^{*[2]}(\Omega)$. We first solve the following problem

$$\min_{\mathcal{P}'^{[2]}_{i,j}} \ \sum\nolimits_{k=1}^{2^{R_{i,j}}} \sum\nolimits_{V^{*}(\mathsf{o}_i(t))\in P'^{[2]}_{i,j,k}} \Big| V^{*[2]}\big(\mathsf{o}_i(t)\big) - \mu^{''}_{k} \Big|. \tag{6}$$

Afterwards, as shown in Figure 1, we cluster the observation any points $\mathsf{o}_i, \mathsf{o}'_i \in \Omega$ based on their corresponding values $V^{*[2]}(\mathsf{o}_i), V^{*[2]}(\mathsf{o}'_i)$, and not based on their original values. Accordingly, each agent i solves the problem (6) for $N_i \leq N$ number of times where, N^{c}_{i} stands for the distinct number of bit-budgets at which agent i has to communicate with other agents in the network.

C. Decentralized Training Phase

After obtaining the communication policies, we solve the following distributed control design problem

$$\underset{\pi_i^m}{\text{argmax}} \quad \mathbb{E}_{\pi_i}\Big\{\mathbf{g}(t')\Big\}, \quad \forall i \in \mathcal{N} \tag{7}$$

through a distributed training phase to obtain the control policy of each agent i, where the expectation is taken over the MAS's trajectory that is influenced by both the control policy $\pi_i^c(\cdot)$ and the communication/quantization policy $\pi_i^m(\cdot)$ of all agents $i \in \mathcal{N}$. The detailed recipe of ESAIC can be found in Algorithm 1 and its performance will be studied both analytically and numerically in sections IV and V, respectively.

As will be shown in section IV, the number of agents in the training phase can be reduced, regardless of the specific method used to compute the function $V^{*[2]}(\cdot)$. Accordingly, we conjecture that other similar (deep) reinforcement learning algorithms can be used for a two-agent centralized training phase to approximate the value function $V^{*[2]}(\cdot)$ - as long as the condition of theorem 2 is met.

Algorithm 1 : ESAIC

Input: γ, α, c
Initialize all-zero Q-table $Q_i^m(\cdot) \leftarrow Q_i^{m,(k-1)}(\cdot)$, for $i = 1:N$
  and all-zero Q-table $Q\big(\mathsf{s}(t), \mathsf{m}(t)\big)$.
Obtain $\pi^{*[2]}(\cdot)$ & $Q^{*[2]}(\cdot)$ by solving (4) using Q-learning.
Compute $V^{*[2]}\big(\mathsf{o}_i(t)\big)$ following eq. (43) in [7], for $\forall \mathsf{o}_i(t) \in \Omega$.
Obtain $\pi_i^{c[2]}$ by solving the problem (5) N_i^c times, for $i = 1:N$.
for each episode $k = 1:K$ **do**
  Randomly initialize the observation $\mathsf{o}_i(t=1)$, for $i = 1:N$
  for $t_k = 1:M$ **do**
    Select $\mathsf{c}_i(t)$ following $\pi_i^{c[2]}(\cdot)$, for $i = 1:N$
    Obtain message $\tilde{\mathsf{c}}_i(t)$, for $i = 1:N$
    Update $Q_i^m\big(\mathsf{o}_i(t-1), \tilde{\mathsf{c}}_i(t-1), \mathsf{m}_i(t-1)\big)$, for $i = 1:N$
    Select $\mathsf{m}_i(t) \in \mathcal{M}$ and follow ϵ-greedy, for $i = 1:N$
    Obtain reward $r\big(\mathsf{s}(t), \mathsf{m}(t)\big)$, for $i = 1:N$
    Make a local observation $\mathsf{o}_i(t)$, for $i = 1:N$
    $t_k = t_k + 1$
  end
  Compute $\sum_{t=1}^{M} \gamma^{t-1} r_t$ for the lth episode
  update ϵ via: $\epsilon = -0.99k/K + 1$
end
Output: $Q_i^m(\cdot)$ and $\pi_i^m\big(\mathsf{m}_i(t)|\mathsf{o}_i(t), \tilde{\mathsf{c}}_i(t)\big)$, for $i = 1:N$

IV. Analytical study of ESAIC

This section provides analytical studies on the average return performance as well as the computational complexity of the proposed ESAIC. Due to the space limit the proof of theorem 2 is removed and is present in the extended version of the manuscript [14].

A. Average return performance

The main result of this subsection is to prove that by solving the problem (5), one can obtain inter-agent communication/quantization policies which are as effective as the solutions to the problem (13) in [7]. Equivalently, one can reduce the number of agents in the centralized training phase and yet draw enough insights from it to design task-oriented communication policies. The proof provided in this section, therefore, is a testament to how rich is the value function of a two-agent centralized training phase to indirectly incorporate the features of the control task into the task-oriented communication design problem (5).

Theorem 2. *Let the bijection $f(\cdot) : \mathcal{V}^{[2]} \rightarrow \mathcal{V}^{[N]}$ be the mapping from the value of observations for a two-agent scenario to the N-agent. For all $i, j \in \mathcal{N}$, the partition $\mathcal{P}_{i,j}^{[2]}$ proposed by ESAIC (that is obtained by solving the problem (5)) are the same as the partition $\mathcal{P}_{i,j}^{[N]}$ proposed by SAIC (that is obtained by solving the problem (13) in [7]) if*

$$c_1: \quad \forall k \in \{1, \dots, B_{i,j}\} \ \exists k' \in \{1, \dots, B_{i,j}\}: \quad \ddot{f}(\mathcal{P}'^{[2]}_{i,j,k}) = \mathcal{P}'^{[N]}_{i,j,k'}. \tag{8}$$

Proof. The is removed due to the space limit and is available in the extended version of this manuscript [14]. ■

Remark 1: Following the theorem 2, all the guarantees that are presented for the performance of SAIC are in place if $R_{i,j} = R \ \ \forall i, j \in \mathcal{N}$.

B. Computational complexity

As is discussed in [15], the computational complexity of exact Q-learning is proportional to the size of state-action space. Exact Q-learning is used in the centralized and distributed training phases of SAIC and ESAIC. In the centralized training phase of SAIC, the computational complexity $\mathcal{O}(|\Omega \times \mathcal{M}|^N)$ grows exponentially with the size of MAS N. Accordingly, the addition of each agent to the system multiplies the complexity of the Q-learning by $|\Omega \times \mathcal{M}|$. The complexity $\mathcal{O}(|\Omega \times \mathcal{M}|^2)$ of the centralized training phase in ESAIC with respect to the size of the MAS N, however, is constant time. That is, ESAIC will always execute at the same time (or space) regardless of the size of the MAS N.

The complexity $\mathcal{O}(|\Omega \times \mathcal{C}^{n-1} \times \mathcal{M}|)$ of the Q-learning problem that each agent solves in SAIC, at the decentralized training phase, also grows exponentially with the addition of each agent to the system. Compared with the centralized training phase, in the distributed training phase, SAIC is much less sensitive to the addition of an agent to the system. Although the complexity of the Q-learning at each agent i multiplies by a constant $|\mathcal{C}|$ with the addition of each agent to the system, the size of the communication space $|\mathcal{C}|$ is much smaller than $|\Omega \times \mathcal{M}|^2$. In the decentralized training phase, ESAIC follows the same complexity patterns.

[2]To understand why "the size of the communication space $|\mathcal{C}|$ is much smaller than $|\Omega \times \mathcal{M}|$", remember that we solve the problem (5) to significantly reduce the size of the communication message space $\mathcal{C}$ of agent i compared with the size of its observation space Ω.

Remark: If the condition c_1 of theorem 2 is met, ESAIC offers the same performance as SAIC at a much reduced computational cost. Accordingly, for a problem comprised of N agents, the time complexity of SAIC is $|\Omega \times \mathcal{M}|^{N-2}$ times higher than ESAIC.

V. Numerical Studies

To evaluate our proposed method, ESAIC, in this section, we leverage numerical experiments on a specific cooperative task i.e., a geometric consensus problem with finite observability, called the rendezvous problem. Geometric consensus problems are emerging in many new applications, such as UAV/vehicle platooning, which makes them a useful application domain for the framework proposed in this paper [12].

A rendezvous problem is a geometrical consensus problem where the goal of a team of agents $\mathcal{N}$ is to simultaneously arrive at a goal point $\omega^T \in \Omega$, while each agent i is only aware of its own location $\mathsf{o}_i(t) \in \Omega$. We consider a square $p \times p$ grid to be where agents move and operate and for it to be the observation space $\Omega = \{0, 1, ..., p^2 - 1\}$ of all agents. As soon as the goal point is visited by one agent (or more) an episode terminates leading to non-deterministic time-horizons T'. Accordingly, all state realizations which correspond to the termination of an episode are illustrated by $\mathcal{S}^T = \{\langle \mathsf{o}_1(t), ..., \mathsf{o}_n(t)\rangle \in \mathcal{S} \,|\, \exists i \in \mathcal{N} : \mathsf{o}_i(t) \in \omega^T\}$.

We also define the subset $\mathcal{S}^T_{n'} \subset \mathcal{S}^T$ that includes all the terminal states where only n' agents have simultaneously reached the goal point i.e.,

$$\mathcal{S}^T_{n'} = \{\langle \mathsf{o}_1(t), ..., \mathsf{o}_n(t)\rangle \in \mathcal{S} \,|\, \forall i \in \mathcal{N}' : \mathsf{o}_i(t) \in \omega^T\},$$

with $\mathcal{N}' \subseteq \mathcal{N}$ being a subset of all agents and $n' = |\mathcal{N}'|$. Accordingly, the subset $\mathcal{S}^T_N \subset \mathcal{S}^{\mathcal{T}}$ is a collection of all terminal states in which all agents have reached the goal location. At the initial time step $t = 1$, the location of all agents is randomly selected amongst the non-goal locations, i.e., for each agent $i \in \mathcal{N}$ the initial position is selected following a uniform distribution $\mathbf{o}_i(1) \sim \mathcal{U}\{\Omega - \{\omega^T\}\}$.

Accordingly, given observations $\langle \mathsf{o}_i(t+1), ..., \mathsf{o}_n(t+1)\rangle$ and actions $\langle \mathsf{m}_1(t+1), ..., \mathsf{m}_n(t+1)\rangle$, all agents receive a single team reward

$$r\big(\mathsf{o}_1(t), ..., \mathsf{o}_n(t), \mathsf{m}_1(t), ..., \mathsf{m}_n(t)\big) = \begin{cases} C_1, & \text{if } P_1 \\ C_2, & \text{if } P_2, \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

where $C_1 < C_2$ and the propositions P_1 and P_2 are defined as $P_1 : T\big(\mathsf{o}_1(t), ..., \mathsf{o}_n(t), \mathsf{m}_1(t), ..., \mathsf{m}_n(t)\big) \in \mathcal{S}^T - \mathcal{S}^T_n$ and $P_2 : T\big(\mathsf{o}_1(t), ..., \mathsf{o}_n(t), \mathsf{m}_1(t), ..., \mathsf{m}_n(t)\big) \in \mathcal{S}^T_N$. When only a subset $\mathcal{N}'$, $|\mathcal{N}| = n' < N$ of agents arrives at the target point ω^T, an episode is ended with the small reward C_1 being obtained, and the large team-reward $C_2 >> C_1$ is accrued when every agent $i \in \mathcal{N}$ simultaneously arrives at the goal location. Naturally, this reward function demands further coordination between agents which in turn can encourage agents to achieve effective communication among themselves.

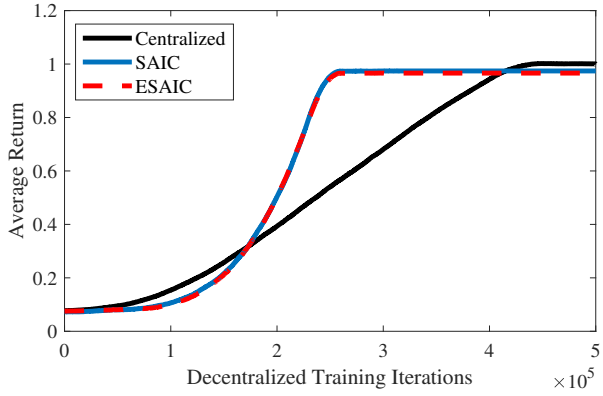

Figure 2. Comparison of the obtained average return via SAIC and ESAIC in MAS in the decentralized training phase while the condition c_1 in (8) is violated.

Moreover, each agent i follows its communication policy at every time step t to obtain a communication $\mathbf{c}_{i,j}(t) \in \mathcal{C} = \{0, 1\}^{R_{i,j}}$ to transmit to every other agent $j \in \mathcal{N}_{-i}$, where $R_{i,j}$ (bits per channel use / per time step) is the fixed bit-budget of agent i when communicating with j. By solving the D-JCCD problem (3), we aim at maximizing the average return of the MAS.

A. Results

ESAIC, SAIC, and centralized schemes are compared by their average return in Fig. 2. The figure is intended to show the applicability of the ESAIC scheme in more complex geometric consensus environments. The size of the grid world for this figure is 8×8, and the multi-agent system is composed of three agents. The figure demonstrates that the performance of ESAIC closely follows that of SAIC, with almost similar average return performance as well as the speed of convergence. The centralized scheme, which is represented by the solid black curve, achieves optimal performance but requires virtually twice the time required for the convergence of ESAIC and SAIC. Fig. 2 suggests that ESAIC is a promising approach for achieving high average return performance in complex MASs, with similar performance to SAIC and faster convergence time than the centralized scheme.

As discussed earlier in section IV, SAIC suffers from prohibitively high computational complexity in its centralized training phase. ESAIC is introduced in this paper to tackle the issue of complexity in the centralized training phase by designing the communication policies only according to a two-agent centralized training. Figure 3 compares the run time required for the implementation of the centralized training phase in both schemes SAIC and ESAIC - both theoretically and analytically. This figure is plotted for the grid worlds of smaller size i.e., 3×3 across all schemes. The analytical results reflect the explanations provided at IV-B.

To realize the end-to-end time required for the training of both algorithms, Fig. 4 is brought. This figure illustrates the

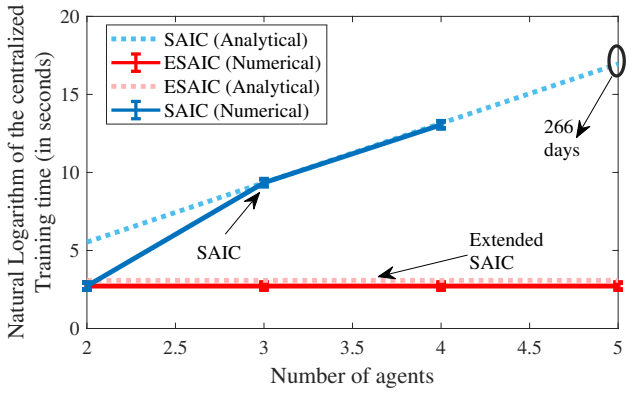

Figure 3. Comparison of the average time required to carry out the centralized training phase in both algorithms SAIC and ESAIC.

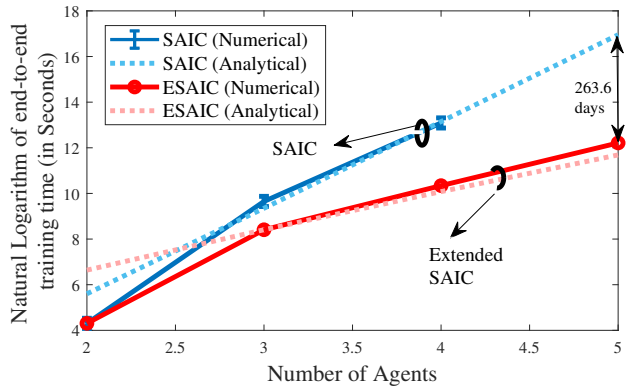

Figure 4. Comparison of the average time required to carry out end-to-end training in both algorithms SAIC and ESAIC.

combined time required to carry out the centralized as well as the decentralized training phase. Inter-agent communications are considered to be $R_{i,j} = 2$ (bits per channel use) across all agents $\forall i, j \in \mathcal{N}$. With an increase in the number of agents, the size of the received communication message space $\mathcal{C}^{n-1}$ increases exponentially leading to an increase in the end-to-end complexity of both algorithms SAIC and ESAIC. Nevertheless, the goal of solving the problem (5) is to significantly reduce the size of each agent's communication transmission space $\mathcal{C}$ compared with the observation space Ω. Accordingly, the exponential increase in the size of received communication space has a much less pronounced impact on the overall complexity of both algorithms. Yet, we expect the size of the received message space $\mathcal{C}^{n-1}$ to be another bottleneck of SAIC that ESAIC can not solve. This bottleneck gets more serious when the number of agents goes double digits. The analytical results reflect the explanations provided at IV-B.

VI. Conclustion

This paper presents a novel and scalable approach for quantifying the value/importance of the observed information in a multi-agent system. In particular, we observe that the computational complexity of obtaining the VoI can be drastically reduced by gaining insights from a similar two-agent MAS - rather than the original N-agent MAS. Yet, the obtained measure to quantify the value of agents' observations is sufficiently rich to help design task-effective multi-agent communication/quantization policies. ESAIC quantized agents' observations such that the observations which are more important/valuable for the cooperative control task are communicated at a higher precision. The result of the analytical as well as numerical studies demonstrates a striking reduction in the end-to-end computational complexity of the communications and control co-design. The proposed algorithm, ESAIC, holds substantial implications for communication system design within multi-agent systems, offering promising applications across various domains including autonomous vehicles, robotics, and wireless sensor networks.

References

[1] J. Shao, Y. Mao, and J. Zhang, "Task-oriented communication for multidevice cooperative edge inference," *IEEE Transactions on Wireless Communications*, vol. 22, no. 1, pp. 73–87, 2023.

[2] X. Kang, B. Song, J. Guo, Z. Qin, and F. R. Yu, "Task-oriented image transmission for scene classification in unmanned aerial systems," *IEEE Transactions on Communications*, vol. 70, no. 8, pp. 5181–5192, 2022.

[3] T.-Y. Tung, S. Kobus, J. P. Roig, and D. Gündüz, "Effective communications: A joint learning and communication framework for multi-agent reinforcement learning over noisy channels," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2590–2603, 2021.

[4] A. Mostaani, O. Simeone, S. Chatzinotas, and B. Ottersten, "Learning-based physical layer communications for multiagent collaboration," in *2019 IEEE Intl. Symp. on Personal, Indoor and Mobile Radio Communications*, Sep. 2019.

[5] P. A. Stavrou and M. Kountouris, "The role of fidelity in goal-oriented semantic communication: A rate distortion approach," *IEEE Transactions on Communications*, pp. 1–1, 2023.

[6] H. Zou, C. Zhang, S. Lasaulce, L. Saludjian, and H. V. Poor, "Goal-oriented quantization: Analysis, design, and application to resource allocation," *IEEE Journal on Selected Areas in Communications*, 2022.

[7] A. Mostaani, T. X. Vu, S. Chatzinotas, and B. Ottersten, "Task-oriented data compression for multi-agent communications over bit-budgeted channels," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 1867–1886, 2022.

[8] F. A. Oliehoek, M. T. Spaan, and N. Vlassis, "Optimal and approximate q-value functions for decentralized pomdps," *Journal of Artificial Intelligence Research*, vol. 32, pp. 289–353, 2008.

[9] D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein, "The complexity of decentralized control of markov decision processes," *Mathematics of operations research*, vol. 27, no. 4, pp. 819–840, 2002.

[10] A. Mostaani, T. X. Vu, S. Chatzinotas, and B. Ottersten, "Task-effective compression of observations for the centralized control of a multi-agent system over bit-budgeted channels," 2023. [Online]. Available: https://arxiv.org/abs/2301.01628

[11] R. A. Howard, "Information value theory," *IEEE Transactions on systems science and cybernetics*, vol. 2, no. 1, pp. 22–26, 1966.

[12] A. Barel, R. Manor, and A. M. Bruckstein, "Come together: Multi-agent geometric consensus," *arXiv preprint arXiv:1902.01455*, 2017.

[13] J. N. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[14] A. Mostaani, T. X. Vu, H. Habibi, S. Chatzinotas, and B. Ottersten, "Task-oriented communication design at scale," *arXiv preprint arXiv:2305.08481*, 2023.

[15] M. G. Azar, R. Munos, M. Ghavamzadaeh, and H. J. Kappen, "Speedy q-learning," 2011.