Task-Oriented Data Compression for Multi-Agent Communications Over Bit-Budgeted Channels

Arsham Mostaani, Thang X. Vu, Symeon Chatzinotas, and Björn Ottersten

***Abstract*—Various applications for inter-machine communications are on the rise. Whether it is for autonomous driving vehicles or the internet of everything, machines are more connected than ever to improve their performance in fulfilling a given task. While in traditional communications the goal has often been to reconstruct the underlying message, under the emerging task-oriented paradigm, the goal of communication is to enable the receiving end to make more informed decisions or more precise estimates/computations. Motivated by these recent developments, in this paper, we perform an indirect design of the communications in a multi-agent system (MAS) in which agents cooperate to maximize the averaged sum of discounted one-stage rewards of a collaborative task. Due to the bit-budgeted communications between the agents, each agent should efficiently represent its local observation and communicate an abstracted version of the observations to improve the collaborative task performance. We first show that this problem can be approximated as a form of data-quantization problem which we call task-oriented data compression (TODC). We then introduce the state-aggregation for information compression algorithm (SAIC) to solve the formulated TODC problem. It is shown that SAIC is able to achieve near-optimal performance in terms of the achieved sum of discounted rewards. The proposed algorithm is applied to a geometric consensus problem and its performance is compared with several benchmarks. Numerical experiments confirm the promise of this indirect design approach for task-oriented multi-agent communications.**

***Index Terms*—Task-oriented communications, semantic communications, data quantization, machine learning for communications, communications for machine learning.**

I. Introduction

The design of traditional communication systems has often been carried out according to task-agnostic principles. Information and coding theories drive the major analytical and design techniques, where the former sets the upper bounds on the system capacity, and the latter focuses on techniques for approaching the bounds with infinitesimal error probabilities. Accordingly, digital communications have made astonishing strides in terms of performance, enabling robust information transmission even under adverse channel conditions. However, in the era of cyber-physical systems, the effectiveness of communications is not solely dictated by the traditional performance indicators (e.g., bit rate, latency, jitter, fairness etc.), but most importantly by the efficient completion of the task in hand, e.g., remotely controlling a robot, automating a production line or collaboratively sensing/communicating through a drone swarm.

The authors are with the Centre for Security Reliability and Trust, University of Luxembourg, Luxembourg. Emails: {arsham.mostaani, thang.vu, symeon.chatzinotas, bjorn.ottersten}@uni.lu

This work is supported by the ERC AGNOSTIC project, ref. H2020/ERC2020POC/957570/DREAM.

Machine to machine communications occur since the received signals can help the receiving end to make more informed decisions or more precise estimates/computations. In this context, the reliability of the communications is not essential beyond serving the specific needs of the control/estimation/computational task that the receiving end machine is trying to accomplish. This calls for a fresh look into the design of communication systems that have been engineered with reliability as one of their ultimate goals. The emerging literature on semantic communications as well as goal/task-oriented communications is trying to take the first steps towards the above-mentioned goal, i.e., incorporating the semantics as well as the goal/usefulness of the message exchange into the design of communication systems [1]–[3]. By jointly analyzing the features of the collaborative task and the constraints on the underlying communication infrastructure, the communication strategies can be adapted or tailored such that they will be specifically effective for the task.

This paper attempts to take the first steps towards designing an *indirect* task-effective data compression theory. While the data compression algorithm proposed by this paper is designed in an *indirect*[1] fashion i.e., not for a specific task, we demonstrate its applicability in a specific task: a geometric consensus problem under finite observability [6]. As attested by [7], "a unified framework to support various tasks is still missing in multi-user semantic communications.". Unlike earlier task-oriented quantization techniques that tailor a quantization scheme to certain application [8], this work proposes an *indirect* design for its task-oriented quantization scheme - SAIC. The *indirect* design is carried out in a fashion that the it never benefits from any explicit domain knowledge about any specific task e.g., geometric consensus problems. Accordingly, the *indirect* design of the algorithms allows them to be applied beyond the geometric consensus problems and to a much wider range of tasks. The framework can be applied where a major communication bottleneck is in place between multiple cooperative decision makers. This bottleneck can occur due to a multitude of reasons (i) the energy lifetime of the communicating agents e.g., in the case of UAV/LEO satellite communications, that forces agents to communicate with low-

[1]By using the word *indirect* here we are not referring to the concept of indirect access to the source of information [4] - this usage of the word falls in the nomenclature of source coding and information theory. In fact, we are referring to the concept being introduced by the control theory nomenclature in which an indirect design is generic enough to be used for an unmodelled system dynamics and not a certain dynamic [5]. Thus the schemes - such as SAIC - which enjoy from an indirect design can be applied to all/a wider range of tasks. In contrast to indirect schemes, "the direct schemes aim at guaranteeing or improving the performance of the cyber-physical system at a particular task by designing a task-tailored communication strategy" [1].

energy high-range communication protocols [9], [10] (ii) the limitations imposed by the environment on the communication channel e.g., in space/underwater missions or (iii) limited communication resources of the network through which agents communicate. For more on the applications of TODC see [1], [11].

A. Task-Oriented Data Compression

In particular, we consider a cooperative scenario where our goal is to optimize the expected return of a multi-agent system that is run on top of an underlying Markov decision process. The system's return is an unknown function of joint observations and control actions of all agents. The system's expected return can be controlled or optimized by selecting the proper joint controls actions at all agents. The partial observability of each agent together with their limitation to merely select local actions necessitates the presence of inter-agent communications to improve the coordination across the multi-agent system. We assume a full mesh communication network between all agents and that all the communication channels in the network are bit-budgeted but error-free. That is, the communication channels are all error-free fixed-rate bit pipes [12] and not variable rate bit-pipes [13] - the fixed rate of communications is constant across all inter-agent communication channels. Under these circumstances, rate-limited communication channels between agents drive the need for task-oriented data compression i.e., the usefulness of each message exchange should be incorporated into the design of the data compression strategy. The communicated messages between agents are useful only when they positively affect the decision-making of the receiving agents towards improving the system's expected return.

The problem we address would be a classic multi-agent Markov decision process (MAMDP) [14] if, each agent's communication message could include all the information inside the agent's observations. We assume, however, that the communication message of each agent is sent over a bit-budgeted communication channel i.e., per each channel use each agent will be able to reliably communicate a bit sequence with a length less than the entropy rate of the observation process. With this information constraint in place, it becomes imperative to carry out the communications at each agent such that they lead to the optimal expected return performance of the MAS. Each agent has to jointly select its control and quantized message at each time step with the aim of optimizing the expected return.

Due to the bit-budgeted communications between the agents, it is necessary for agents to compactly represent their observations in communication messages. As we ultimately measure the performance of the MAS in terms of the expected return, the loss of information caused by the compact representation of the agents' observations needs to be managed in such a way that it minimally affects the obtained return [15], [16]. As such, in this form of compression scheme which we call task-oriented data compression, *the goal of abstraction is different from conventional compression schemes* whose ultimate aim is to reduce the distortion between the original signal and the decoded/reconstructed signal [17] - see [8], [18], where a similar task-based notion is introduced and a comparison of it with our work in Table I.

B. Literature Review

As we study the joint communication and control design of a MAS, the topic of this paper falls under the general category of multi-agent communications [19]. In contrast to many other cooperative multi-agent systems [20], the full state and action information are not available here to each agent. Accordingly, agents are required to carry out communication to overcome these barriers [19]. Earlier works used to address the coordination of multiple agents through a noise-free communication channel, where the agents follow an engineered communication strategy [21]–[25]. Later the impact of stochastic delays in multi-agent communication was considered on the multi-agent coordination [24], while [25] considers event-triggered local communications. Deep reinforcement learning with communication of the gradients of the agents' objective function was proposed in [26] to learn the communication among multiple agents. In contrast to the above-mentioned works, the presence of noise in the inter-agent communication channel was first studied by [27] where exact reinforcement learning was used to design the inter-agent communications. Later, the authors of [16] proposed a deep reinforcement learning approach to address a similar problem. Papers [8], [16], [18], [27], [28] and [29] have contributed to the rapidly emerging literature on task-oriented communications [1]. Noteworthy are also some novel metrics that are introduced in [30] to measure the *positive signaling* and *positive listening* amongst agents which learn how to communicate [26], [27], [29].

The current work can also be seen as designing a state aggregation algorithm. In this paper, state aggregation enables each agent to compactly represent its observations through communication messages while maintaining their performance in the collaborative task. Classical state aggregation algorithms, however, have been used to reduce the complexity of the dynamic programming problems over MDPs [31]–[34] as well as Partially Observable MDPs [35]. One similar work is [36], which studies a task-based quantization problem. In contrast to our work, the assumption there is that the parameter to be quantized is only measurable and cannot be controlled. In our problem, agents' observations stem from a generative process with memory, an MDP. Similarly, in [37], the authors have introduced a gated mechanism so that reinforcement learning-aided agents reduce the rate of their communication by removing messages which are not beneficial for the team. However, their proposed approach mostly relies on numerical experiments. In contrast, this paper relies on analytical studies to design a multi-agent communication policy which efficiently coordinates agents over a bit-budgeted channel - the benefits of our analytical approach are briefly explained in the contributions section I-C. State aggregation algorithms are often developed for single-agent scenarios and are used to reduce the complexity of MDPs. To the best of our knowledge,

we are the first to design a TODC algorithm using state-aggregation schemes. In particular, we use state-aggregation to design a data compression scheme to compactly represent the observation process of each agent in a multi-agent system.

Conventionally, the communication system design is disjoint from the distributed decision-making design [21]–[24], [26], [38]. The current work can also be interpreted as a demonstration of the potential of the joint design of the data compression/quantization and control policies. Determining the existence of a quantizer operating at a certain bit-budget to achieve a given figure of expected return is known to be an "intriguing open problem" [15] - even for single agent scenarios. Here we set a non-closed form upper bound on the expected-return performance of the multi-agent system given a quantization data rate/ the finite size of the discrete alphabet of the quantizer. We show how this joint quantization and control design problem is connected to minimizing an absolute error distortion measure via Theorem 1. A similar interpretation of the TODC problem can also be seen in [39]. While relevant, their setup is different from our work as they consider two distortion criteria for the rate-distortion problem.

We will show in section II-B, that, in fact, the decentralized problem we target can be translated as the joint constrained design of the control policies as well as the observation function of a Dec-POMDP to maximize the expected return. While in classic Dec-POMDP problems the observation function is considered to be a fixed function [40], by a constrained design of the observation function, our problem setting offers more flexibility in designing a multi-agent system. The design of the observation function helps to filter the non-useful observation information of each agent while meeting the problem's constraint i.e., the communication bit-budget. The mathematical framework being used here is neither a classic MDP as we have the issue of partial observability, nor is a partially observable MDP (POMDP) [41] as the action vector is not jointly selected at a single entity. Our problem setting is differentiated from Dec-POMDPs due to the fact that in Dec-POMDPs the partial observability is accepted as is, where as in our problem setting we design the lens through which the agents acquire a partial observation/perception of the environment.

Nevertheless, a similar class of problems - often referred to as task-oriented, goal-oriented or efficient communication approaches, has recently received significant attention from the communication society, see e.g., the extensive surveys on similar problems in [1]–[3]. Table I positions the current work against some of the recent research that is closely related. To date, there is no work in the literature that we are aware of, which provides an analytical approach to the design of task-based communications for the coordination of multiple cooperative agents.

C. Contributions

The contributions of this paper are as follows:

Firstly, we develop a general cooperative multi-agent framework in which agents interact over an underlying MDP environment. Unlike the existing works which assume perfect communication links [26], [29], [38], [42], we assume the practical bit-budgeted communications between the agents. We formulate a multi-agent cooperative problem where agents interact over an underlying MDP and can communicate over a bit-budgeted channel. Our goal is to derive the optimal control and communication strategies to maximize the expected return. We will show in section II-B, that an underlying difference in our setting from the Dec-POMDP is that here we carry out a constrained design of each agent's perception function - which is also referred to as the observation function in the literature of the Dec-POMDP [43]. The constraints of this design are dictated by the bit-budget of the inter-agent communication channels.

Secondly, Theorem 1, in section III, derives the interconnection between the joint control and communication/quantization problem and a generalized version of the data quantization problem: TODC problem. In fact, the TODC problem distils all the relevant features of the control task and takes them into account in a novel non-conventional communication design problem. This is the underlying reason behind the effectiveness of the designed communications and is one the contributions in this work differentiating it from existing works in [8], [15], [16], [18], [26], [27], [30], [44]. Our analytical studies show that how the value function - the function that estimates the expected return of the system given the current observation - can be considered as a proper indirect measure of the usefulness of the data to be compressed. Thus, Theorem 1, shows how the usefulness of the (observation) data can be incorporated into the design of the TODC policy.

Thirdly, we propose a novel algorithm - SAIC - as a multi-agent state-aggregation algorithm which designs indirect task-effective communication strategies via solving (an approximated version of) the TODC problem. As a result, the performance of SAIC in terms of the system's expected return is on par with the jointly optimal strategies. To the best of our knowledge, this is the first use of state-aggregation algorithms for data-compression applications (in multi-agent systems) according to which our work differs from the classic state-aggregation literature [31]–[34] as well as the recent advancements in multi-agent communication literature [26], [30].

Moreover, we extend the existing results in the single-agent state-aggregation literature [33] on the gap between the optimal control and the state-aggregated control schemes, where the former has access to the true state of the environment and the latter has access to an aggregated state of the environment - to reduce the computational complexity. We quantify the same gap for a multi-agent system - Theorem 8. In our work, however, the gap is due to the bit-budget that is introduced on the inter-agent communication channels, whereas in classic state-aggregation literature the gap was a consequence of the constraints on the computational complexity. In addition to that, our theoretical results show that if our proposed method, SAIC, is applied the expected return of the multi-agent communication system - with the bit-budget in place -

Table I
COMPARISON BETWEEN OUR WORK AND THE RELATED PRIOR ART

Paper	Information source with memory	Joint coms and control	Distributed	Source/Channel coding Quantization	Implicit/Explicit coms	Analytical/ Data-driven
[8], [18]	×	×	×	Quantization	N/A	Data-driven
[44]	×	✓	✓	N/A	Implicit	Analytical
[15]	✓ (Linear)	✓	×	Quantization	Explicit	Analytical
[26], [30]	N/A	✓	✓	N/A	Explicit	Data-driven
[16], [27]	✓(Markov)	✓	✓	Channel Coding	Explicit	Data-driven
Our work	✓(Markov)	✓	✓	Quantization	Explicit	Analytical

can stay in close proximity to the optimal expected return that is obtained under jointly optimal strategies.

Last but not least, numerical experiments are carried out on a geometric consensus problem to compare the performance of SAIC with several other benchmark schemes in terms of the optimality of the expected return, for a multi-agent scenario [2]. It is shown that when communication bit-budgets are in place, SAIC is of significant advantage over the benchmarks. In particular, we observe a very tight gap between the performance of SAIC and the optimal control strategy where only the latter runs over perfect communication channels and the former runs over bit-budgeted channels.

D. Organization

Section II describes the system model for a cooperative multi-agent task with rate-constrained inter-agent communications. Section III Proposes a scheme for the joint design of communication and control policies that takes the value of information into account to perform data compression. We also provide analytical results on how distant the result of this algorithm can be from the optimal centralized solution. The numerical results and discussions are provided in section IV. Finally, section V concludes the paper.

E. Notation

For the reader's convenience, a summary of the notation that we follow in this paper is given in Table II. Bold font is used for matrices or scalars which are random and their realizations follows simple font.

II. SYSTEM MODEL

In the multi-agent system, comprised of n agents, at any time step t each agent $i \in \mathcal{N}$ makes a local observation $\mathsf{o}_i(t) \in \Omega$ on environment while the true state of the environment

$$\mathsf{s}(t) = \langle \mathsf{o}_1(t), ..., \mathsf{o}_n(t) \rangle \tag{1}$$

is a member of $\mathcal{S} = \Omega^n$. The alphabets Ω and $\mathcal{S}$ define observation space and state space, respectively. The particular observation structure of agents' observations, is referred to as collective observations in the literature [19]. Under collective observability, individual observation of an agent provides it with partial information about the current state of the environment, however, having knowledge of the collective observations acquired by all of the agents is sufficient to realize the true state of environment - eq. (1). The columns of the state vector are orthogonal to each other. Note that even in the case of collective observability, for agent i to be able to observe the true state of environment at all times, it needs to have access to the observations of the other agents $j \in \mathcal{N} - \{i\} \triangleq \mathcal{N}_{-i}$ through communications at all times.

The true state of the environment $\mathsf{s}(t)$ is controlled by the joint actions $\mathsf{m}(t) = \langle \mathsf{m}_1(t), ..., \mathsf{m}_n(t) \rangle \in \mathcal{M}^n$ of the agents, where each agent i can only choose its local action $\mathsf{m}_i(t) \in \mathcal{M}$. The environment runs on discrete time steps $t = 1, 2, ..., M$, where at each time step, each agent i selects its domain level action $\mathsf{m}_i(t)$ upon having an observation $\mathsf{o}_i(t)$ of the environment. Dynamics of the environment are governed by a conditional probability mass function (CMF)

$$T\big(\mathsf{s}(t+1)|\mathsf{s}(t), \mathsf{m}(t)\big) = p\big(\mathsf{s}(t+1)|\mathsf{s}(t), \mathsf{m}(t)\big) \tag{2}$$

which is unknown to the agents. $T(\cdot) : \Omega^{2n} \times \mathcal{M}^n \to [0, 1]$ determines the future state of the environment $\mathsf{s}(t+1)$ given its current state $\mathsf{s}(t)$ and the joint actions $\mathsf{m}(t)$. We recall that each agent i's domain level action $\mathsf{m}_i(t)$ 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.

A deterministic reward function $r(\cdot) : \Omega^n \times \mathcal{M}^n \to \mathbb{R}$ indicates the reward of all agents at time step t, where the arguments of the reward function are the joint observations $\mathsf{s}(t)$ and the domain-level joint actions $\mathsf{m}(t)$ of all agents. We assume that the underlying environment over which agents interact can be defined in terms of an MDP [3] determined by the tuple $\big\{\Omega^n, \mathcal{M}^n, r(\cdot), \gamma, T(\cdot)\big\}$, where Ω and $\mathcal{M}$ are discrete alphabets, $r(\cdot)$ is a function, $T(\cdot)$ is defined in (2) and 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 [45], a realization of the

[2]Due to the complexity related issues explained in section V & VI, the numerical results are limited to two-agent and three-agent scenarios.

[3]As defined in the literature [10], the underlying MDP' is the horizon-T' MDP defined by a hypothetical single agent that takes joint actions $\mathbf{m}(t) \in \mathcal{M}^n$ and observes the nominal state $\mathbf{s}(t) \triangleq \langle \mathbf{o}_1(t), \ldots, \mathbf{o}_n(t) \rangle$ that has the same transition model $T(\cdot)$ and reward model $R(\cdot)$ as the environment experienced by our multi-agent system.

function $r\big(\mathsf{s}(t),\mathsf{m}(t)\big)$ will be accessible for all agents at some time steps. Since the tuple $\big\{\Omega^n,\mathcal{M}^n,r(\cdot),\gamma,T(\cdot)\big\}$ is an MDP and the state process $\mathbf{s}(t)$ is jointly observable by agents, the system model of this cooperative multi-agent setting, under perfect communications, is also referred to as a multi-agent MDP (MAMDP or MMDP) in the literature of multi-agent decision making [14], [46], [47].

In what follows two problems regarding the above-mentioned setup is detailed i.e., centralized and decentralized control problems. The main intention of this paper is to address decentralized control which also incorporates inter-agent communications for a system of multiple agents. The centralized control problem, however, is also formalized in subsection II-A as the optimal expected return obtained for the centralized problem can serve as a lower-bound/(upper-bound) for the decentralized scheme. Moreover, the simpler nature and mathematical notations used for the centralized problem, allow the reader to have a smoother transition to the decentralized problem which is of a more complex nature.

A. Centralized Control

We consider a scenario in which a central controller has instant access to the observations $\mathsf{o}_1(t),...,\mathsf{o}_n(t)$ of both agents through a free (with no cost on the objective function) and reliable communication channel. From the central controller's point of view, the environment is the same as the underlying MDP that governs the system $\Big\{\Omega^n,\mathcal{M}^n,r(\cdot),\gamma,T(\cdot)\Big\}$. The goal of the centralized controller is to maximize the expected sum of discounted rewards (3). The expectation is computed over the joint PMF of the whole system trajectory $\mathbf{s}(1),\mathbf{m}(1),...,\mathbf{s}(M),\mathbf{m}(M)$ from time $t=1$ to $t=M$, where this joint probability mass function (PMF) is generated if agents follow policy $\pi(\cdot)$, eq. (4), for their action selections at all times and the initial state $\mathbf{s}(1)\in\mathcal{S}$ is randomly selected by the initial distribution $\mathbf{s}(1)\sim\alpha_{\mathbf{s}}$. For the sake of having 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 $\langle\mathsf{o}_1(t),...,\mathsf{o}_n(t),\mathsf{m}_1(t),...,\mathsf{m}_n(t)\rangle$ and the realization of the whole system trajectory by $\{\mathsf{tr}(t)\}_{t=1}^{t=M}$. Accordingly, the problem boils down to a single agent problem which can be denoted by

$$\max_{\pi(\cdot)}\qquad \mathbb{E}_{p_\pi\left(\{\mathsf{tr}(t)\}_{t=1}^{t=M}\right)}\Big\{\sum_{t=1}^{M}\gamma^{t-1}r\big(\mathbf{s}(t),\mathbf{m}(t)\big)\Big\}\tag{3}$$

where the policy π can be expressed as a CMF

$$\pi\Big(\mathsf{m}(t)\Big|\mathsf{s}(t)\Big)=p\Big(\mathsf{m}(t)\Big|\mathsf{s}(t)\Big),\tag{4}$$

and $p_\pi\big(\mathsf{s}(t+1)|\mathsf{s}(t)\big)$ is the probability of transitioning from $\mathsf{s}(t)$ to $\mathsf{s}(t+1)$ when the joint action policy $\pi(\cdot)$ is executed by the central controller. Similarly, $p_\pi\big(\{\mathsf{tr}(t)\}_{t=1}^{t=M}\big)$ is the joint PMF of $\mathsf{tr}(1),\mathsf{tr}(2),...,\mathsf{tr}(M)$ when the joint action policy $\pi(\cdot)$ is followed by the central controller.

On one hand, problem (3) can be solved using single-agent Q-learning [45] and the solution $\pi^*(\cdot)$ obtained by Q-learning is guaranteed to be the optimal control policy, given some non-restricting conditions [48]. On the other hand, the use-cases of the centralized approach are limited to the applications in which there is a permanent communication link with an unlimited bit-budget between the agents and the controller. Whereas these conditions are not met in many remote applications, where there is no communication infrastructure to connect the agents to the central controller.

Given sufficient training time, and channels with the sufficient rate of communication between the agents and the central controller, the centralized algorithm provides us with a performance upper bound in maximizing the objective function (3). Perfect communication between the central controller and distributed agents, however, may not exist due to the resource limitations of the telecommunication/communication network. Thus, the aim of this paper is to introduce decentralized approaches which are run over practical bit-budgeted communication channels, yet show comparable performance levels. In the distributed scenario, the agents do not communicate with a central controller, but the bit-budgeted communications are performed for inter-agent message exchange. The centralized problem can be presented by an MDP and be solved efficiently by a single agent reinforcement learning algorithm. As explained in the section I-C, the decentralized problem is a more complicated/general form of Dec-POMDP, where we know that a Dec-POMDP is already much more complex than an MDP to solve [43] - to see further insights about the significance and the applications of the decentralized problem see e.g., [1].

Table II
TABLE OF NOTATIONS

Symbol	Meaning
$\mathbf{x}(t)$	A generic random variable generated at time t
$\mathsf{x}(t)$	Realization of $\mathbf{x}(t)$
$\mathcal{X}$	Alphabet of $\mathbf{x}$(t)
$\lvert\mathcal{X}\rvert$	Cardinality of $\mathcal{X}$
$p_{\mathbf{x}}\big(\mathsf{x}(t)\big)$	Shorthand for $\Pr\big(\mathbf{x}(t)=\mathsf{x}(t)\big)$
$H\big(\mathbf{x}(t)\big)$	Information entropy of $\mathbf{x}(t)$ (bits)
$\mathcal{X}_{-\mathbf{x}}$	$\mathcal{X}-\{\mathbf{x}\}$
$\mathbb{E}_{p(\mathsf{x})}\{\mathbf{x}\}$	Expectation of the random variable X over the probability distribution $p(\mathsf{x})$
$\delta(\cdot)$	Dirac delta function
$\mathsf{tr}(t)$	Realization of the system's trajectory at time t

B. Problem Statement

Here we consider a scenario in which the same objective function explained in Eq. (3) needs to be maximized by the multi-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, and not indirect (implicit) communications [44], [49]. However, the communication is done through a bit-budgeted but reliable channel. The bit-budget of the channel is R-bits per time step. Equivalently, each agent i at every time step t produces and transmits a single digit communication message $\mathbf{c}_i(t)\in\mathcal{C}$ such that

$$log_2|\mathcal{C}|\leq R,\tag{5}$$

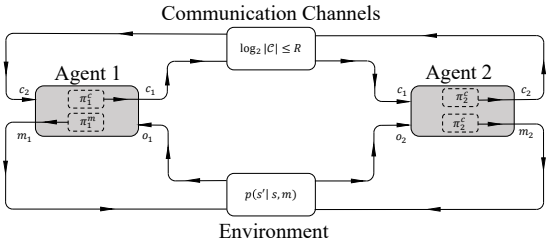

Figure 1. An illustration of the decentralized cooperative two-agent system with rate-limited inter-agent communications.

i.e., the size of the code-books $\mathcal{C}$ for all agents is the same and is less than 2^R. The communication message $\mathbf{c}_i(t)$ produced by agent i is broadcast and received every agent $j \in \mathcal{N}_{-i}$. It should be noted that the design of the channel coding is beyond the scope of this paper and the main focus is on the compression of agents' observations. In particular we consider R to be time-invariant and to follow:

$$R < \min\left\{H\big(\mathbf{o}_1(t)\big), ..., H\big(\mathbf{o}_n(t)\big)\right\}. \tag{6}$$

The above-mentioned information constraint which will be in place throughout this paper together with the observation structure assumed in eq. (1) are of the aspects that distinguish our work from many of the related works in the literature of multi-agent communications [16], [27]. Now let the function $\mathbf{g}(t^{'})$ denote the system's *return*:

$$\mathbf{g}(t^{'}) = \sum\nolimits_{t=t^{'}}^{M} \gamma^{t-1} r\big(\mathbf{s}(t), \mathbf{m}(t)\big). \tag{7}$$

Note that $\mathbf{g}(t^{'})$ is a 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 the arguments of this function. In contrast to the centralized problem, the goal of the decentralized problem is to jointly design the communication/quantization as well as $\pi_i^c(\cdot)$ control policies $\pi_i^m(\cdot)$ for each agent $i \in \mathcal{N}$ to maximize the average return of the system. The control policy $\pi_i^m : \mathcal{M} \times \mathcal{C}^{n-1} \times \Omega \to [0,1]$ of each agent i is defined as CMF

$$\begin{aligned}&\pi_i^m\Big(\mathsf{m}_i(t)\Big|\mathsf{o}_i(t), \mathsf{c}_{-i}(t)\Big) = \\ &\Pr\Big(\mathbf{m}_i(t) = \mathsf{m}_i(t)\Big|\mathbf{o}_i(t) = \mathsf{o}_i(t), \mathbf{c}_{-i}(t) = \mathsf{c}_{-i}(t)\Big),\end{aligned} \tag{8}$$

in which, $\mathsf{c}_{-i}(t) \in \mathcal{C}^{n-1}$ is a vector that includes all communication messages $\mathsf{c}_j(t)$, $\forall j \in \mathcal{N}_{-i}$. The communication policy $\pi_i^c : \Omega \times \mathcal{C}^{n-1} \to \mathcal{C}$ of each agent i is a deterministic data quantization (many to one) function:

$$\mathsf{c}_i(t) = \pi_i^c\Big(\mathsf{o}_i(t), \mathsf{c}_{-i}(t)\Big), \tag{9}$$

which has a discrete domain $\Omega \times \mathcal{C}$, making the quantizer a discrete quantizer. The joint control policy π^m is a tuple made of n elements with its i-th element being $\pi_i^m(\cdot)$. Similarly, The joint communication policy π^c is another tuple with its i-th element being $\pi_i^c(\cdot)$.

According to the above definitions, the decentralized joint control and communication design problem is formalized as

$$\begin{aligned}\max_{\pi_i^m, \pi_i^c} \quad & \mathbb{E}_{p_{\pi^m,\pi^c}\left(\{\mathsf{tr}(t)\}_{t=1}^{t=M}\right)}\Big\{\mathbf{g}(1)\Big\}, \quad i \in \mathcal{N} \\ \text{s.t.} \quad & log_2|\mathcal{C}| \le R,\end{aligned} \tag{10}$$

where the expectation is taken over $p_{\pi^m,\pi^c}\big(\{\mathsf{tr}(t)\}_{t=1}^{t=M}\big)$ which is the joint PMF of $\mathsf{tr}(1), \mathsf{tr}(2), ..., \mathsf{tr}(M)$ when each agent $i \in \mathcal{N}$ follows the action policy $\pi_i^m(\cdot)$ and the communication policy $\pi_i^c(\cdot)$ and the initial state $\mathbf{s}(1) \in \mathcal{S}$ is randomly selected by the initial distribution $\mathbf{s}(1) \sim \alpha_{\mathbf{s}}$. Given communication policy $\pi_i^c(\cdot)$, $\forall i \in \mathcal{N}$, we now define the perception function $h_i(\cdot) : \mathcal{S} \to \mathcal{C}^{n-1} \times \Omega$ of agent i which is the lens through which agent i perceives the state $\mathbf{s}(t)$ of the environment.

$$\begin{aligned}&h_i\big(\mathbf{s}(t)\big) = \\ &\langle \pi_1^c(\mathbf{o}_1(t)), \pi_2^c(\mathbf{o}_2(t)), ..., \mathbf{o}_i(t), \pi_{i+1}^c(\mathbf{o}_{i+1}(t)), ..., \pi_n^c(\mathbf{o}_n(t))\rangle\end{aligned} \tag{11}$$

Agent i's perception of the environment is characterized by the communication policy $\pi_j^c(\cdot)$ of each agent $j \in \mathcal{N}_{-i}$. Accordingly, agent i uses its sensory signal $\mathbf{o}_i(t)$ together with the received communication signals $\mathbf{c}_{-i}(t)$ to acquire its perception of the environment. While the perception function defined here plays a role very similar to the observation function in Dec-POMDPs [40], the main difference is that here we design communication policies such that they directly affect the perception of agents from the environment. In contrast, in the case of Dec-POMDPs, the observation function is given. Communication policies $\pi_j^c(\cdot), \forall j \in \mathcal{N}_{-i}$ partially define the perception function of agent i.

To make the problem more concrete, further to (8) and (9), here we assume the presence of instantaneous and synchronous communications between agents, contrasting with the delayed [27], [50] and sequential communication models. Fig. 2 demonstrates this communication model during a single time-step. As such, each agent i at any time step t prior to the selection of its action $\mathsf{m}_i(t)$ receives the communication vector $\mathsf{c}_{-i}(t)$ that encodes the observations of each agent $j \in \mathcal{N}_{-i}$ at time t.

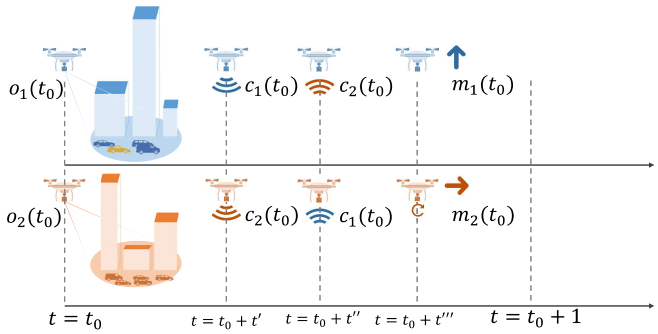

Figure 2. Ordering of observation, communications and action selection for synchronous and instantaneous communication model in a multi-UAV object tracking example, with $0 < t' < t'' < t''' < 1$. At time $t = t_0$ both agents (UAVs) make local observations on the environment. At time $t = t_0 + t'$ both agents select a communication signal to be generated. At time $t = t_0 + t''$ agents receive a communication signal from the other agent. At time $t = t_0 + t'''$ agents select a domain level action, here it can be the movement of UAVs or rotation of their cameras etc.

In a general approach, the selection of communication action $\mathsf{c}_i(t)$ at agent i could be conditioned on both $\mathsf{o}_i(t)$ and $\mathsf{c}_{-i}(t)$. Since we assume instantaneous and synchronous inter-agent communications, here we are focused on communication policies of type $\pi_i^c\big(\mathsf{o}_i(t)\big)$, where communication actions of each agent at each time are selected only based on its observation at that time. For clear reasons, it is not possible to adopt a synchronous and instantaneous inter-agent communication model and yet take the communication message $\mathsf{c}_{-i}(t)$ into account when selecting the communication $\mathsf{c}_i(t)$ at agent i.

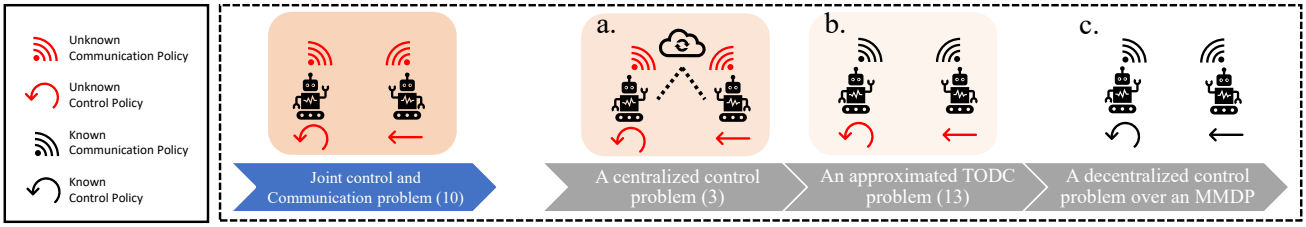

Figure 3. Here we show how we approached solving the joint control and communication problem for a distributed multi-agent system in a sequence of steps. According to the legend, one can understand that at the end of each step what are the known and unknown policies. a. This step solves the problem (3) for a centralized multi-agent system where the objective is to design one centralized control strategy. b. This step solves the problem (13) for a distributed multi-agent system where the objective is to design the communication policies of all agents. c. this step solves the problem for a distributed multi-agent system where the objective is to design the control policies of all agents.

Here we assume that the communication resources are split evenly amongst the agents, by considering the bit-budget of all communication channels to be equal to R. As such, each agent $i \in \mathcal{N}$ encodes its observation $\mathsf{o}_i(t)$ to $\mathsf{c}_i(t)$ using a code-book $\mathcal{C}$ of the same length $|\mathcal{C}|$ - with the constraint (5) in place.

III. State Aggregation for Information Compression (SAIC) in Multi-Agent Coordination Tasks

The main result of this section - provided by Theorem 1 - is to show that finding the quantization policy in the joint control and quantization problem (10) can be approximated by a TODC problem. The goal of this problem is to quantize the observations of all agents according to how valuable these observations are within any specific task. The value of observations should be measured by the value function $V^*(\cdot)$ - eq. (25). Lemma 2 approximates the TODC to a k-median clustering of the of observations according to their values, while lemma 4 computes the value function of each agent's observation. The concluding remarks of this section study the convergence and the optimality of the decentralized control policies.

Fig. 3 is brought to demonstrate the chronological order according to which a joint communication and quantization is solved by SAIC. Our proposed scheme, SAIC, breaks down the joint communication and quantization problem to smaller problems that are feasible to solve. In this section, however, the subsections are organized according to the logical order that these smaller problems are encountered: (A) in section III-A , we address the communication design of multi-agent communications by transforming the primary joint control and quantization problem (10) to a novel problem (12) called TODC - step "b" of the Fig. 3. (B) Since solving the TODC problem relies on the knowledge of the value function $V^*(\cdot)$, it is necessary to obtain the value function $V^*(\cdot)$ prior to solving the TODC problem. In section III-B, the optimal value function $V^*(\cdot)$ is obtained via a centralized training phase - step "a" of the Fig. 3. Given the knowledge of the value function $V^*(\cdot)$, the TODC problem incorporates the features of the specific control task in the communication design problem. Accordingly, we can separately solve the communication problem with very little compromise on the optimality of the system's expected return. (C) As the final step, in section III-C, decentralized training phase is carried out to distributively design the control policy of each agent given the communication/quantization policy obtained via solving the TODC problem. Decentralized training is shown in step "c" of the Fig. 3. Since we follow standard methods to carry out the centralized training - steps "a" of the Fig. 3 - we will be mainly focused on deriving and solving the TODC problem and providing guarantees on the performance of the MAS in the decentralized training phase - steps "b" and "c" of the Fig. 3 respectively. Fig. 4 illustrates how SAIC performs data compression while it maintains the performance of the multi-agent system in its task.

A. Task-Oriented Data Compression Problem

The main result of this section is provided by Theorem 1. This theorem departs from the joint communication/quantization and control problem and arrives at the task-oriented data compression problem (12).

Theorem 1. *The design of the communication policy in problem (10) can be approximated as a generalized data quantization problem*

$$\begin{aligned} \underset{\pi_i^c(\cdot)}{min} \quad & \mathbb{E}_{p_{\pi^m,\pi^c}\left(h_i\left(\mathbf{s}(1)\right)\right)}\left\{\left|V^*\left(\mathbf{s}(1)\right) - V^*\left(h_i\left(\mathbf{s}(1)\right)\right)\right|\right\} \\ s.t. \quad & log_2|\mathcal{C}| \leq R, \end{aligned} \tag{12}$$

in which the measure of distortion is the absolute difference of the value functions $V^*\left(\mathbf{s}(t)\right)$ and $V^*\left(h_i\left(\mathbf{s}(1)\right)\right)$ with the source of information $\mathbf{s}(t) \in \Omega^n$ being a Markovian stochastic process. The function $V^*\left(h_i\left(\mathbf{s}(1)\right)\right)$ measures the optimal value of the perceived state $h_i\left(\mathbf{s}(1)\right)$ from agent i's perspective.

Proof. Appendix A. ■

In Appendix A-C, we provide more details on how to obtain the value $V^*\left(h_i\left(\mathbf{s}(1)\right)\right)$ of the perceived state from the agent i's point of via Lemma 12. This value function allows us to indirectly quantify the usefulness of agent i's observation. With this interpretation in mind, in the TODC problem (12), unlike conventional quantization problems, we are not minimizing the absolute difference between the original signal $\mathbf{s}(1)$ and its quantized version $h_i\left(\mathbf{s}(1)\right)$. Instead, we

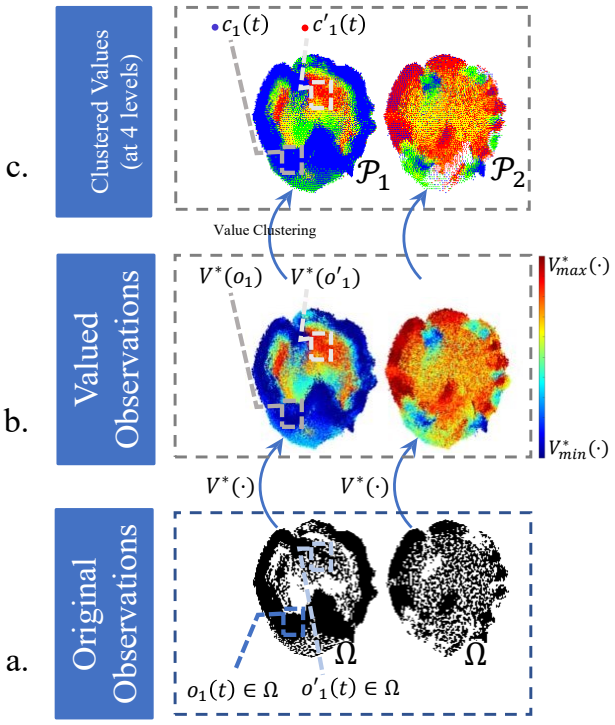

Figure 4. The subplots of this figure illustrate how in SAIC we transform a high-dimensional (σ-dimensions) and high-precision observation space into aggregated one-dimensional low-precision/digitized communication message space. This figure is plotted for a scenario where $R = 2$ (bits per channel use) and thus, observation values are clustered at $2^R = 4$ different levels. a. A 2D demonstration of the original high-dimensional and high-precision observation space of agents is shown here in black and white. b. After carrying out the centralized training phase we will obtain the value function $V^*(\cdot)$ - which acts as indirect measure of the usefulness of observation data to be communicated. Now by applying the value function $V^*(\cdot)$ at every point of the original observation space we get valued observations - a one-dimensional high-precision space as the output space of the value function $V^*(\cdot)$. c. By clustering the observation points according to their corresponding values for each agent i we would get a one-dimension and low-precision/digitized communication message space. The quantization illustrated in this diagram is using only 4 levels of quantization that are represented by 4 colours. All the points in the observation space of the agent i which are represented by the same colour, in subplot c, will be represented by a unique communication message - i.e., the accuracy of the original data is reduced and hence requires fewer communication bits to be transmitted. Accordingly, agent 1, after observing $\mathsf{o}_1(t)$ transmits the communication message $\mathsf{c}_1(t)$ which is a compressed version of $\mathsf{o}_1(t)$ while it maintains the performance of the multi-agent team in maximizing their expected return.

are minimizing the distance between how useful/valuable the original signal $\mathbf{s}(1)$ is and how useful the quantized version of the signal $h_i(\mathbf{s}(1))$ are for the task at hand. This is in-line with what many believe as the mission of the goal-oriented/task-oriented communications. Let us recall that the value function here is an *indirect* measure of usefulness, as it can be obtained for any task that can be expressed via Markov Decision Processes - making it a measure of usefulness that is applicable to a plethora of scenarios [1], [11].

The significance of the result obtained by Theorem 1 is multi-fold: (i) Multi-dimensional observations will be transformed to one-dimensional output space of the value functions, reducing the complexity of the clustering algorithm, (ii) It can be shown that the observation points will be linearly separable when being clustered according to the problem (12), (iii) It is widely accepted that the mission of goal oriented communications is to incorporate the usefulness/value of the data for the task when designing the task-effective communications. The result of Theorem 1, in which the design of the quantizer relies on the value/usefulness of observations resonates well with this purpose of goal-oriented communications. (iv) It is known that the value of observations starts to grow as we get closer to the ultimate target of the task in hand. With this interpretation of "target" in mind, the finding of Theorem 1 is in line with the adaptive quantization schemes, which stretch the quantization intervals when the observations are far from the target and sharpen the quantization when the observations are closer to the target [13], [51]. This interpretation is also confirmed by our numerical results in section V, Fig. 8.

To solve a quantization problem as (12) using non-variational techniques, it is customary to approximate/convert a quantization problem by/to a clustering problem [52], [53]. Lemma 2 approximates the quantization problem (12) by a clustering problem.

Lemma 2. *The quantization problem (12) can be approximated by a clustering problem*

$$\min_{\mathcal{P}_i} \quad \sum\nolimits_{k=1}^{|\mathcal{C}|} \sum\nolimits_{\mathsf{o}_i(t)\in\mathcal{P}_{i,k}} \left| V^*\big(\mathsf{o}_i(t)\big) - \mu_k^{'} \right|, \tag{13}$$

where μ'_k is the centroid of the k-th cluster $\mathcal{P}_{i,k}$ and $\mathcal{P}_i = \{\mathcal{P}_{i,1}, \ldots, \mathcal{P}_{i,|\mathcal{C}|}\}$ is a partition of the observation space Ω. Similar to any other quantization function, the quantizer $\pi_i^c(\cdot)$, can be uniquely described by the partition $\mathcal{P}_i$ together with $\mathcal{C}$.

Proof. Appendix B provides proof and discussions. ■

The problem (13), can be solved via k-median clustering. In order to that, one can first perform the k-median clustering on the observation values by solving

$$\min_{\mathcal{V}_i} \quad \sum\nolimits_{k=1}^{2^B} \sum\nolimits_{V^*(\mathsf{o}_i(t))\in\mathcal{V}_{i,k}} \left| V^*\big(\mathsf{o}_i(t)\big) - \mu_k^{''} \right|,$$

where $\mathcal{V}_i$ is the set of all observation values of agent i and $\{\mathcal{V}_{i,1}, ..., \mathcal{V}_{i,|\mathcal{C}|}\}$ is its partition. Afterwards, as shown in Figure 4, the observation points should be clustered according to the clustering of their corresponding values. That is, any two distinct observation points $\mathsf{o}'_i, \mathsf{o}''_i \in \Omega$ are clustered together in $\mathcal{P}_{i,j}$ if and only if their values $V^*(\mathsf{o}'_i), V^*(\mathsf{o}''_i) \in \mathcal{V}_{i,j}$ are in the same cluster $\mathcal{V}_{i,j}$.

Theorem 1 together with lemma 2 allows us to find a communication/quantization policy $\pi_i^c(\cdot)$ by clustering the input space Ω of the communication policy according to the values $V^*\big(\mathsf{o}_i(t)\big)$ of the input points. The performance guarantees for the obtained communication/quantization policy will be shown in section IV. One can obtain $V^*\big(\mathsf{o}_i(t)\big)$ via solving the centralized problem (3) by Q-learning. The subsection III-B, details a centralized training approach for obtaining the value observations $V^*\big(\mathsf{o}_i(t)\big)$.

B. Centralized Training Phase

While solving the TODC problem can provide us with a task-effective design of quantization policy, to solve (12) we need to know the value of observations according to the optimal centralized control policy. By solving the centralized problem (3), the value of joint observations and actions $Q^*\big(\mathsf{s}(t), \mathsf{m}(t)\big)$ can be obtained. Let us recall that the centralized training phase will only yield an optimal policy if the environment is jointly observable - as described by condition 3.

Condition 3.

$$\mathsf{s}(t) = \langle \mathsf{o}_1(t), ..., \mathsf{o}_n(t) \rangle. \tag{14}$$

Accordingly, following the lemma 4 we can compute the value of each agent's observations $V^*\big(\mathsf{o}_i(1)\big)$. But before lemma 4, let us first give an intuitive/philosophical meaning of the centralized training and distributed execution. We know that in task-oriented communication design, our goal is to take into account the usefulness/value of the data for the task in hand. Thus we need to first be able to measure the usefulness/value of the data to be transmitted. The centralized training phase is needed to come up with a precise measure of usefulness for the specific task in hand. We have already shown in Theorem 1, that this measure of usefulness is nothing but the value observations $V^*\big(\mathsf{o}_i(1)\big)$ - yet the exact function values can be known only after the centralized training phase. During the centralized training phase, we assume perfect communication between all agents and a central controller - this is a common practice in the literature of multi-agent communications and coordination [26], [54]. Whereas, in the decentralized training - step "c" of the Fig. 3 - as well as in the execution phase, we assume bit-budgeted communications. That is, all the results reported for SAIC in section V are obtained via bit-budgeted communications.

Lemma 4. *One can compute the $V^*\big(\mathsf{o}_i(1)\big)$ following*

$$V^*\big(\mathsf{o}_i(t)\big) = \sum_{\mathsf{o}_{-i}(t) \in \Omega^{n-1}} \max_{\mathsf{m}} \; Q^*\big(\mathsf{s}(t), \mathsf{m}(t)\big) p\big(\mathbf{o}_{-i}(t) = \mathsf{o}_{-i}(t)\big).$$

Proof. Appendix C. ∎

Based on (15), $V^*\big(\mathsf{o}_i(1)\big)$ can be computed both analytically (if transition probabilities of environment are available) and numerically. As detailed in Algorithm 1, SAIC first solves a centralized control problem to compute the value $V^*(\mathsf{o})$ for all $\mathsf{o} \in \Omega$ - this is equivalent to the step "a" of the Fig. 3 and subplot (b) of the Fig. 4. Afterwards, SAIC solves the approximated TODC problem (12) by converting it to a k-median clustering (13), leading to an observation aggregation/quantization function for each agent i determined by $\pi_i^c(\cdot)$ - this is equivalent to the step "b" of the Fig. 3 and the subplot (c) of the Fig. 4. By following this aggregation function, the observations $\mathsf{o}_i(t) \in \Omega$ will be aggregated/quantized such that the performance of the multi-agent system in terms of the objective function it attains is optimized. As SAIC uses a deterministic mapping of observation o_i to produce the communication message c_i, SAIC is guaranteed to have positive signalling [30].

C. Obtaining Decentralized Control Policies via a Decentralized Training Phase

Upon the availability of the $\pi_i^c(\cdot),\ \forall i \in \mathcal{N}$, which was obtained by solving problem (13), we need to find control policies for all agents corresponding to the communication policies $\pi_i^c(\cdot),\ i \in \mathcal{N}$. That is, we now solve the problem (10) by plugging the exact communication policy $\pi_i^c(\cdot)\ \forall i \in \mathcal{N}$ into it. Within this training phase - referred to as the decentralized training phase - control Q-tables $Q_i^m(\cdot)\ \forall i \in \mathcal{N}$ are obtained - step "c" of the Fig. 3. This training phase, as well as the execution phase of the algorithm, can both be carried out distributively, while agents communicate over bit-budgeted channels using the communication policies obtained before in section III-A. The following remarks are brought to characterize the performance of SAIC, in the decentralized training phase.

We now first define the concept of lumpablity, according to which we will then set a condition - Condition 6 - for the correctness of remarks 3 and 4.

Definition 5. ***Lumpability of an MDP:*** *Let $\alpha_{\mathbf{s}}$ be the probability distribution of the initial state of an MDP at the initial step. The MDP is called (strongly) lumpable with respect to the perception function $h_i(\cdot)$ if the the transitions between all the perceived states $h_i(\mathbf{s}(t))$ - which are perceived through the lens of $h_i(\cdot)$ - follow Markov rule for every probability distribution $\alpha_{\mathbf{s}}$ of the initial state of the original MDP [34].*

Condition 6. *Let the environment as perceived from the perspective of agent i within the decentralized training phase be called an aggregated MDP denoted by $\Big\{\Omega \times \mathcal{C}^{n-1}, \mathcal{M}, r(\cdot), \gamma, T'(\cdot)\Big\}$, whereas the state space of the aggregated MDP $\Omega \times \mathcal{C}^{n-1}$ is an image of Ω^n under the perception function $h_i(\cdot)$. Now given the definition 5, assuming the lumpability of the underlying MDP $\Big\{\Omega^n, \mathcal{M}^n, r(\cdot), \gamma, T(\cdot)\Big\}$ with respect to $h_i(\cdot)$ is equivalent to the assumption that the aggregated $\Big\{\Omega \times \mathcal{C}^{n-1}, \mathcal{M}, r(\cdot), \gamma, T(\cdot)\Big\}$ is an MDP under every possible $\alpha_{\mathbf{s}}$. This assumption is in place for the correctness of remarks 3 and 4.*

Remark 1: The optimal policy $\pi^*(\cdot)$ is achievable by the centralized training phase. Assuming Condition 3 to hold, the environment is fully observable for the central controller while the central controller posses the ability to jointly select the actions for all agents. The problem will thus reduce to a single agent Q-learning applied on an MDP with asymptotic convergence to the optimal policy $\pi^*(\cdot)$.

Remark 2: During the decentralized training phase, each agent, instead of viewing the environment as the original

Algorithm 1 State Aggregation for Information Compression (SAIC)

Input: γ, α, c
Initialize all-zero table $N_i^m\big(\mathsf{o}_i(t), \mathsf{c}_{-i}(t), \mathsf{m}_i(t)\big)$, for $i \in \mathcal{N}$
  and Q-table $Q_i^m(\cdot) \leftarrow Q_i^{m,(k-1)}(\cdot)$, for $i \in \mathcal{N}$
  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 (3) using Q-learning [45].
Compute $V^*\big(\mathsf{o}_i(t)\big)$ following eq. (15), for $\forall \mathsf{o}_i(t) \in \Omega$.
Solve problem (13) by applying k-median clustering to obtain $\pi_i^c(\cdot)$, for $i \in \mathcal{N}$.
for each episode $k = 1 : K$ **do**
  Randomly initialize local observation $\mathsf{o}_i(t=1)$, for $i \in \mathcal{N}$
  for $t_k = 1 : M$ **do**
    Select $\mathsf{c}_i(t)$ following $\pi_i^c(\cdot)$, for $i \in \mathcal{N}$
    Obtain message $\mathsf{c}_{-i}(t)$, for $i \in \mathcal{N}$
    Update $Q_i^m\big(\mathsf{o}_i(t-1), \mathsf{c}_{-i}(t-1), \mathsf{m}_i(t-1)\big)$, for $i \in \mathcal{N}$
    Select $\mathsf{m}_i(t) \in \mathcal{M}$ following UCB, for $i \in \mathcal{N}$
    Increment $N_i^m\big(\mathsf{o}_i(t), \mathsf{c}_{-i}(t), \mathsf{m}_i(t)\big)$, for $i \in \mathcal{N}$
    Obtain reward $r\big(\mathsf{s}(t), \mathsf{m}(t)\big)$, for $i \in \mathcal{N}$
    Make a local observation $\mathsf{o}_i(t)$, for $i \in \mathcal{N}$
    $t_k = t_k + 1$
  end
  Compute $\sum_{t=1}^{M} \gamma^t r_t$ for the lth episode
end
Output: $Q_i^m(\cdot)$,
  and $\pi_i^m\big(\mathsf{m}_i(t)|\mathsf{o}_i(t), \mathsf{c}_{-i}(t)\big)$ by following greedy policy for $i \in \mathcal{N}$

underlying MDP denoted by $\Big\{\Omega^n, \mathcal{M}^n, r(\cdot), \gamma, T'(\cdot)\Big\}$, views an aggregated form of the original MDP denoted by $\Big\{\Omega \times \mathcal{C}^{n-1}, \mathcal{M}, r(\cdot), \gamma, T'(\cdot)\Big\}$. The aggregated MDP experienced by agent i will be an MDP itself, if the conditions 3 and 6 hold.

Remark 3: The MAS, during the decentralized training phase, will be composed of n different MDPs with identical state space $\Omega \times \mathcal{C}^{n-1}$, action space $\mathcal{M}$ and reward signal. The resulting multi-agent environment will be, according to the definition, a multi-agent MDP (MMDP) [47].

Remark 4: Within the distributed training phase, distributed Q-learning is applied to a deterministic MMDP [4], which leads to an asymptotically optimal control policy [14] [5]. For this remark to be true conditions 3 and 6 must hold.

Note that the control policy $\pi_i^{m,SAIC}(\cdot)$ that is obtained within the distributed training phase of SAIC is optimal for the given communication policy $\pi^{c,SAIC}(\cdot)$, that was obtained within the centralized training phase. Therefore, $\pi_i^{m,SAIC}(\cdot)$ is not necessarily an optimal solution to the problem (10). In Theorem section IV, however, we set an upper-bound on the possible loss on the expected return of the system due to the joint selection of $\pi_i^{m,SAIC}(\cdot)$ and $\pi^{c,SAIC}(\cdot)$.

[4]The definition of MMDP in [47] is identical to the definition of cooperative MAMDP used in [14].

[5]This training phase can result in an asymptotically optimal control policy of all agents for non-deterministic MMDPs. This, however, will require n additional centralized training phases prior to the decentralized training phase, where n is the number of agents.

IV. Characterizing the error bound of SAIC

As discussed in section III, SAIC uses two approximations to solve the original joint quantization and control problem. It was not, however, explained that how these approximation would impact the performance of SAIC in terms of the system's average return. By extending the results of [33] to a multi-agent scenario, we characterize the performance gap of SAIC proposed in section III. Instead of measuring the difference between the average return obtained by SAIC with that of the jointly optimal policies for the problem (10), in Theorem 8, we measure the performance gap between the average return attained by SAIC with that of the centralized controller - whereas the latter has had access to perfect communications and as well as full observability of the environment. The measured gap is, indeed, larger than the performance gap between SAIC and a hypothetical jointly optimal solution to (10), as in the case of the central controller there is no communication/observation limitation in place. The performance gap between SAIC and the centralized solution provided by Theorem 8 is proposed in terms of the discount factor λ of the task and a positive scalar ϵ. Definition 7 details the notion of ϵ-cost uniform. Lemma 9 is proposed to compute the value of ϵ for SAIC.

Definition 7. *Given a positive number ϵ a subset $\mathcal{P}_{i,k} \subset \Omega$ is said to be ϵ-cost-uniform with respect to the policy $\pi(\cdot)$ if the following conditions hold for two arbitrary observations $\mathsf{o}', \mathsf{o}'' \in \mathcal{P}_{i,k}$:*

$$c_1: \quad \mathcal{M}_\pi(\mathsf{o}') = \mathcal{M}_\pi(\mathsf{o}'') \tag{15}$$

$$c_2: \quad \textit{For any } \mathsf{m} \in \mathcal{M}_\pi(\mathsf{o}'): |Q^\pi(\mathsf{o}', \mathsf{m}) - Q^\pi(\mathsf{o}'', \mathsf{m})| < \epsilon, \tag{16}$$

where $\mathcal{M}_\pi(\mathsf{o}') = \big\{\mathsf{m} \in \mathcal{M} : \pi(\mathsf{m}|\mathsf{o}') > 0\big\}$.

Theorem 8. *Consider a multi-agent system in which agents are subject to local observability and local action selection. If agents are allowed to communicate through communication channels with a bit-budget R-bits at each time step, the maximum achievable expected return of the multi-agent system following SAIC algorithm will be in a small neighbourhood of the same MAS if it was controlled with a centralized unit under perfect communications:*

$$\mathbb{E}_{p_{\pi^*}(\{\mathsf{tr}(t)\}_{t=t_0}^{t=M})}\big\{\mathbf{g}(t_0)\big\} - \mathbb{E}_{p_{\pi_i^m, \pi_i^c}(\{\mathsf{tr}(t)\}_{t=t_0}^{t=M})}\big\{\mathbf{g}(t_0)\big\} < \frac{2\,\epsilon}{(1-\gamma)^2}, \tag{17}$$

where γ is the discount factor and ϵ should be computed according to lemma 9, conditioned on the lumpability of the original MDP - Condition 6.

Proof. Appendix D. ■

In Theorem 8, we will show that the error gap between

Lemma 9. *Given the partition $\mathcal{P}_i = \{\mathcal{P}_{i,1}, ..., \mathcal{P}_{i,2^R}\}$ that is obtained by solving eq. (38) during the centralized training phase, all subsets $\mathcal{P}_{i,k}$ for $k \in \{1, 2, ..., 2^R\}$ are ϵ-cost-*

uniform with respect to the optimal joint policy $\pi^*(\cdot)$ where ϵ can be obtained by the following

$$\epsilon/2 = \max_{k,\mathsf{o}_i} \left| V^*\big(\mathsf{o}_i(t)\big) - \mu_k^{'} \right|. \tag{18}$$

Proof. Following definition 7 and eq. (13) the proof is straightforward. ■

V. Performance Evaluation

In this section, we evaluate our proposed schemes via numerical results for a particular geometric consensus problem with finite observability called the rendezvous problem. Geometric consensus problems arise in numerous emerging applications such as UAV/vehicle platooning - making them a meaningful application area for the framework proposed by this paper [6]. The numerical results achieved by SAIC will prove the suitability of the proposed framework as a potential enabling technology for vehicle/UAV platooning under limited communications.

The rendezvous problem, which is a sub-category of the geometric consensus, has been previously investigated in the literature [42], [55], whereas in our case the inter-agent communication channel is set to have a limited bit-budget. The rendezvous problem is of particular interest to us, also because it allows us to consider a cooperative MAS comprising of multiple agents that are required to communicate for their coordination task. In particular, as detailed in subsection V-A, if the communication between agents is not efficient, at any time step t each agent i will only have access to its local observation $\mathsf{o}_i(t)$, which is its own location in the case of rendezvous problem. This mere information is insufficient for an agent to attain the larger reward C_2, but is sufficient to attain the smaller reward C_1. Accordingly, compared with cases in which no communication between agents is present, in the set up of the rendezvous problem, efficient communication policies can increase the attained objective function of the MAS up to six-folds, as will be seen in Fig. 4. The system operates in discrete time, with agents taking actions and communicating in each time step $t = 1, 2, \dots$. We consider a variety of grid worlds with different size values N and different locations for the goal-point ω^T. We compare the proposed SAIC and LBIC with (i) the centralized Q-learning scheme and (ii) the Conventional Information Compression (CIC) scheme which is explained in subsection V-B. Changing the reward function can also build new scenarios. For example, a reward function that encourages the agents to come together as close as possible but not collide with each other can emulate a vehicle platooning scenario. While useful, it is outside the scope of our work to investigate the response of the multi-agent system to different rewarding schemes. Note that, according to Theorem 1, regardless of the definition of the reward function, the geometric consensus problem (or in general the joint quantization and control problem) can be solved by SAIC if the necessary Conditions 3 and 6 are met, and centralized training phase is feasible. As the number of agents n increases, the Q-learning for the centralized training phase becomes increasingly demanding in terms of computational complexity; this is where SAIC's bottleneck lies.

A. Rendezvous Problem

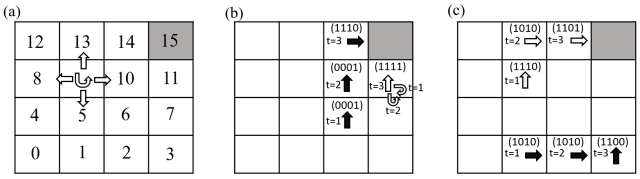

Figure 5. The rendezvous problem when $n = 2$, $N = 4$ and $\omega^T = 15$: (a) illustration of the observation space, Ω, i.e., the location on the grid, and the environment action space $\mathcal{M}$, denoted by arrows, and of the goal state ω^T, marked with gray background; (b) demonstration of a sampled episode, where arrows show the environment actions taken by the agents (empty arrows: actions of agent 1, solid arrows: actions of agent 2) and the $B = 4$ bits represent the message sent by each agent. A larger reward $C_2 > C_1$ is given to both agents when they enter the goal point at the same time, as in the example; (c) in contrast, C_1 is the reward accrued by agents when only one agent enters the goal position [27].

As illustrated in Fig. 5, in a rendezvous problem, multiple agents operate on an $N \times N$ grid world and aim at arriving at the same time at the goal point on the grid. Each agent $i \in \mathcal{N}$ at any time step t can only observe its own location $\mathsf{o}_i(t) \in \Omega$ on the grid, where the observation space is $\Omega = \{0, 1, ..., n^2 - 1\}$. Each episode terminates as soon as an agent or more visit the goal point which is denoted as $\omega^T \in \Omega$. That is, at any time step t that the observation of each agent $i \in \mathcal{N}$ is a member of Ω^T, the episode will be terminated - so the time horizon M is non-deterministic. The subset $\mathcal{S}^T \subset \mathcal{S}$ also defines all state realizations where one or more agents are in the goal location i.e.,

$$\mathcal{S}^T = \{\langle \mathsf{o}_1(t), ..., \mathsf{o}_n(t)\rangle \in \mathcal{S} \,|\, \exists i \in \mathcal{N} : \mathsf{o}_i(t) \in \omega^T\}.$$

We also define the subset $\mathcal{S}^T_{n'} \subset \mathcal{S}^T$ that includes all the terminal states where only n' number of agents have arrived at the goal location i.e.,

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

where $\mathcal{N}' \subseteq \mathcal{N}$ is a subset of all agents with size $|\mathcal{N}'| = n'$. Following the same definition for $\mathcal{S}^T_{n'}$, the subset $\mathcal{S}^T_n$ is equivalent to the set of all terminal states where all agents are at the goal location. At time $t = 1$, the initial position of all agents is randomly and uniformly selected amongst the non-goal states, i.e., for each agent $i \in \mathcal{N}$ the initial position of the agent is $\mathsf{o}_i(1) \in \Omega - \{\omega^T\}$.

At any time step $t = 1, 2, ...$ each agent i observes its position, or environment state, and acquires information about the position of the other agents by receiving a communication message vector $\mathsf{c}_{-i}(t)$ sent by the other agents $j \in \mathcal{N}_{-i}$ at the time step t. Based on this information, agent i selects its environment action $\mathsf{m}_i(t)$ from the set $\mathcal{M} = \{\text{Right}, \text{Left}, \text{Up}, \text{Down}, \text{Stop}\}$, where an action $\mathsf{m}_i(t) \in \mathcal{M}$ represent the horizontal/vertical move of agent i on the grid at time step t. For instance, if an agent i is on a grid-world as depicted on Fig. 5 (a), and observes $\mathsf{o}_i(t) = 4$ and selects "Up" as its action, the agent's observation at the next time step will be $\mathsf{o}_i(t+1) = 8$. If the position to which the agent should be moved is outside the grid, the environment is assumed to keep the agent in its current position. We assume that all these deterministic state transitions are captured by $T\big(\mathsf{o}_1(t), ..., \mathsf{o}_n(t), \mathsf{m}_1(t), ..., \mathsf{m}_n(t)\big)$, which can determine the

observations of agents in the next time step $t+1$ following

$$\langle \mathsf{o}_1(t+1), ..., \mathsf{o}_n(t+1)\rangle = T\big(\mathsf{o}_1(t), ..., \mathsf{o}_n(t), \mathsf{m}_1(t), ..., \mathsf{m}_n(t)\big).$$

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

$$r\big(\mathsf{o}_1(t), ..., \mathsf{o}_n(t), \mathsf{m}_1(t), ..., \mathsf{m}_n(t)\big) = \begin{cases} C_1, & \text{if } P_1 \\ C_2, & \text{if } P_2, \\ 0, & \text{otherwise,} \end{cases} \tag{19}$$

where $C_1 < C_2$ and the propositions P_1 and P_2 are defined as $P_1 : T\big(\mathsf{o}_1(t), ..., \mathsf{o}_n(t), \mathsf{m}_1(t), ..., \mathsf{m}_n(t)\big) \in \mathcal{S}^T - \mathcal{S}_n^T$ and $P_2 : T\big(\mathsf{o}_1(t), ..., \mathsf{o}_n(t), \mathsf{m}_1(t), ..., \mathsf{m}_n(t)\big) \in \mathcal{S}_n^T$. When only a subset $\mathcal{N}'$, $|\mathcal{N}| = n' < n$ of agent arrives at the target point ω^T, the episode will be terminated with the smaller reward C_1 being obtained, while the larger reward C_2 is attained only when all agents visit the goal point at the same time. Note that this reward signal encourages coordination between agents which in turn can benefit from inter-agent communications.

Furthermore, at each time step t agents choose a communication message to send to the other agent by selecting a communication action $\mathsf{c}_i(t) \in \mathcal{C} = \{0,1\}^R$ of R bits, where R (bits per channel use / per time step) is the fixed bit-budget of all inter-agent communication channels. The goal of the MAS is to maximize the average return by solving the problem (10).

B. Conventional Information Compression In multi-agent Coordination Tasks

As a baseline, we consider a conventional scheme that selects communications and actions separately. For communication, each agent i sends its observation $\mathsf{o}_i(t)$ to the other agents by following policy $\pi_i^c(\cdot)$. According to this policy the agent's observation $\mathsf{o}_i(t)$ will be mapped to a binary bit sequence $\mathsf{c}_i(t)$, using an injective (and not necessarily surjective) mapping $f_1 : \Omega \to \{0,1\}^R$. Consequently, the communication policy π_i^c becomes deterministic and follows

$$\pi_i^c\big(\mathsf{c}_i(t+1)|\mathsf{o}_i(t)\big) = \delta\Big(\mathsf{c}_i(t+1) - f_1\big(\mathsf{o}_i(t)\big)\Big). \tag{20}$$

Agent i obtains an estimate $\mathsf{c}_j(t)$ of the observation of all agents $j \in \mathcal{N}_{-i}$ by having access to a quantized version of $\mathsf{o}_j(t)$. This estimate is used to define the environment state-action value function $Q_j^m\big(\mathsf{o}_i(t), \mathsf{c}_{-i}(t), \mathsf{m}_i(t)\big)$. This function is updated using Q-learning and the UCB policy in a manner similar to Algorithm 1, with no communication policy to be learned.

This communication strategy is proven to be optimal [19], if the inter-agent communication does not impose any cost on the cooperative objective function, the communication channel is noise-free and the bit-budget of communication channels are larger than the entropy rate of the observation process $R \geq H(\mathbf{o}_i)$. Under these conditions, and when the dynamics of the environment are deterministic, each agent i can distributively learn the optimal policy $\pi_i^m(\cdot)$, using value iteration or its model-free variants e.g., Q-learning [14]. While this communication policy is optimal only with a channel bit-budget $R \geq H(\mathbf{o}_j)$, in this paper, we are focused on the scenarios with $R \leq H(\mathbf{o}_j)$. Therefore, due to the bit-budget of the communication channel, a form of TODC is required.

Note that compression before a converged action policy is not possible, since all observations are a priori equally likely. Thus, we first train the CIC on a communication channel with unlimited capacity. Afterwards, when a probability distribution for observations is obtained, by applying Lloyd's algorithm [52], we define an equivalence relation on the observation space Ω with 2^R numbers of equivalence classes $\mathcal{Q}_1, ..., \mathcal{Q}_{2^R}$. According to the defined equivalence relation by Lloyd's algorithm, we can uniquely define the mapping $f_1 : \Omega \to \{0,1\}^R$ that maps each agent i's observation $\mathsf{o}_i(t)$ to a communication message $\mathsf{c}_i(t)$. The inverse $f_1^{-1}(\cdot)$ of the quantization mapping that maps agent j's quantized observation $\mathsf{c}_j(t)$ into a estimated observation is not an injective mapping anymore. That is, by receiving the communication message $\mathsf{c}_j(t) \in \mathcal{Q}_k \subset \mathcal{C}$ agent i can not retrieve $\mathsf{o}_j(t)$ but understands the observation of agent j has been a member of $\mathcal{Q}_k$. Note that CIC algorithm has a limitation, as it requires the first round of training to be done over communication channels with unlimited capacity.

C. Results

To perform our numerical experiments, rewards of the rendezvous problem are selected as $C_1 = 1$ and $C_2 = 10$, 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 10% of the experiment iterations. We will use the terms "value of the collaborative objective function", "value of the objective function" and "average return" interchangeably throughout this section. Regardless of the grid-world's size and goal location, the grids are numbered row-wise starting from the left-bottom as shown in Fig. 5-a. Apart from Fig. 7 that illustrates the result related to a rendezvous problem for a three-agent system, other figures have been obtained when experimenting in a two-agent environment. Fig. 6 illustrates the performance of the proposed SAIC as well as six other benchmark schemes

- Centralized Q-learning under perfect communications.
- Learning based information compression (LBIC) is a different indirect scheme to design task-oriented communications which performs the joint design of communication and control policies through reinforcement learning following an algorithm similar to the one proposed in [27].
- CIC, see the details of CIC in subsection V-B.
- Heuristic non-communicative (HNC) algorithm is a direct heuristic scheme which exploits the domain knowledge of its designer about the rendezvous task - making it not applicable to any other task rather than the rendezvous problem. The domain knowledge is utilized to design a control policy where no communication is present. In HNC, agents approach the goal point and wait next to it for a large enough number of time-steps to make sure the other agent has also arrived there. Only after that, they will get into the goal point. Note that this scheme requires

communication/coordination between agents prior to the starting point of the task.

- Heuristic optimal communication (HOC) algorithm is a direct heuristic scheme which exploits the domain knowledge of its designer about the rendezvous task - making it not applicable to any other task rather than the rendezvous problem. The domain knowledge is utilized to design jointly optimal communication and control policies. In HNC, agents approach the goal point and wait next to it until they hear from the other agent it also has arrived there. Only after that, they will get into the goal point. Note that this scheme requires communication/coordination between agents prior to the starting point of the task.
- Hybrid scheme uses the abstract representation of agents' observations according to SAIC with $R = 2$ bits and feeds these latent observations to a centralized controller. The central controller learns the joint action selection of both agents using Q-learning.

It is imperative to recall that, not all the schemes evaluated by Fig. 6 are benefit from indirect designs - making them not sufficiently general to be applied to all other multi-agent communication problems with rate-limited inter-agent channels. Regardless of their effectiveness, SAIC, LBIC, CIC and Hybrid are indirect schemes potentially applicable to any other task-oriented compression problem. Whereas, HNC and HOC are tailor-made for the rendezvous problem. In other words, the knowledge that we have about the rendezvous task is already embedded in HNC and HOC to enable the most effective communication/control strategies. HNC and HOC, however, allow us to understand how effective other indirect approaches are even when no knowledge about the specific rendezvous task is embedded in them.

The performance is measured in terms of the expected sum of discounted rewards in a rendezvous problem. The grid-world is considered to be of size $N = 8$ and its goal location to be $\omega^T = 22$. The bit-budget of the channel between the two agents is $R = 2$ bits per time step. Since centralized Q-learning is not affected by the limitation on the channel's bit-budget, it achieves optimal performance after sufficient training, 160k iterations. The CIC, due to the insufficient bit-budget of the communication channel, never achieves the optimal solution. The LBIC, however, is seen to outperform the CIC, although it is trained and executed fully distributedly. While enjoying a fast convergence, it is observed that the SAIC can achieve optimal performance by less than 1% gap, whereas the performance gap for the LBIC and CIC are much more pronounced ranging from 20% to 30%. The yellow curve showing the performance of the CIC with no communication between agents would show us the best performance of distributed reinforcement learning that can be achieved if no communication between agents is in place without having any domain knowledge - that is present in the HOC and HNC. In fact, the better performance of any scheme compared with the yellow curve, is the sign that the scheme is either benefiting from some effective communication between agents or from some domain knowledge. Note that, when inter-agent communication is unavailable, i.e., $R = 0$ bit per time step, there would be no difference in the performance of the CIC, SAIC or LBIC as all of them use the same algorithm to find out the action policy $\pi_i^m(\cdot)$. We also recall the fact that both the CIC and SAIC require a separate training phase which is not captured by Fig. 5. SAIC requires a centralized training phase - to perform the computations demonstrated in line 5 of the algorithm 1 - and CIC a distributed training phase with unlimited capacity of inter-agent communication channels. The performance of these two algorithms in Fig. 5 is plotted after the first phase of training.

Similar to Fig. 6, the performance of SAIC is illustrated in Fig. 7, this time in a $n = 3$ three-agent system. In this case, the grid-world is considered to be of size $N = 3$ and its goal location to be $\omega^T = 9$. The bit-budget of the inter-agent communication channels is set to be $R = 1$ bits per time step. The shaded area around the curve corresponding to SAIC, shows the standard deviation of SAIC in the training as well as the execution phases - at any given training episode k the width of the shaded curve is equal to the standard deviation of SAIC's return from the training episode k to the episode $k-1000$. This figure illustrates the very robust performance of SAIC in a three-agent scenario. For this particular experiment we used decaying epsilon greedy policies with the starting value of $\epsilon = 1$ and the ending value of $\epsilon = 0.03$. To overcome the issue of credit assignment in multi-agent systems - see e.g., [54] to get familiar with the concept, here we used a different reward function via which we trained the agents. 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

$$r\big(\mathsf{o}_1(t), ..., \mathsf{o}_n(t), \mathsf{m}_1(t), ..., \mathsf{m}_n(t)\big) = \begin{cases} C_2^{n'-1}, & \text{if } P_3, \\ 0, & \text{otherwise,} \end{cases} \tag{21}$$

where the proposition P_3 is defined as $P_3 : T\big(\mathsf{o}_1(t), ..., \mathsf{o}_n(t), \mathsf{m}_1(t), ..., \mathsf{m}_n(t)\big) \in \mathcal{S}_n^{T'}$. When a subset $\mathcal{N}'$, $|\mathcal{N}| = n' \leq n$ of agent arrives at the target point ω^T, the episode will be terminated with the reward $C_2^{n'-1}$ being obtained, while the largest reward C_2^{n-1} is attained only when all agents visit the goal point at the same time. Note that this reward signal encourages coordination between agents which in turn can benefit from inter-agent communications.

To explain the underlying reasons for the remarkable performance of the SAIC, Fig. 8 is provided so that equivalence classes $\{\mathcal{P}_{i,1}, ..., \mathcal{P}_{i,2^R}\}$ computed by the SAIC can be seen - all the locations of the grid shaded with the same colour belongs to the same ϵ-cost-uniform equivalence class. The SAIC is extremely efficient in performing state aggregation such that the loss of observation information barely incurs any loss on the achievable sum of discounted rewards - also depicted in Fig. 5. The Fig. 8-(a), illustrates the state aggregation adopted by the SAIC, for which the average return is illustrated in Fig. 4. It is illustrated in Fig. 8-(a) that how the SAIC performs observation compression with ratio $R_c = 3 : 1$, while it leads to nearly no performance loss for the collaborative task of the MAS. Here the definition of compression ratio follows $R_c = \big\lceil H\big(\mathbf{o}_i(t)\big)\big\rceil / \big\lceil H\big(\mathbf{c}_i(t)\big)\big\rceil$. It was observed in 8 that the observation clusters identified by SAIC have not been linearly separable under their original representation. In contrast, when clustered according to

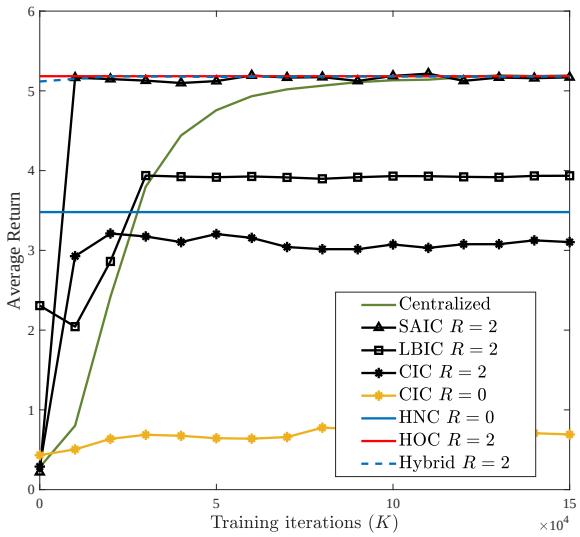

Figure 6. A comparison between all seven schemes in terms of the achievable objective function with the bit-budget of $R = 2$ bits per channel use/time steps and number of training iterations/episodes $K = 200k$.

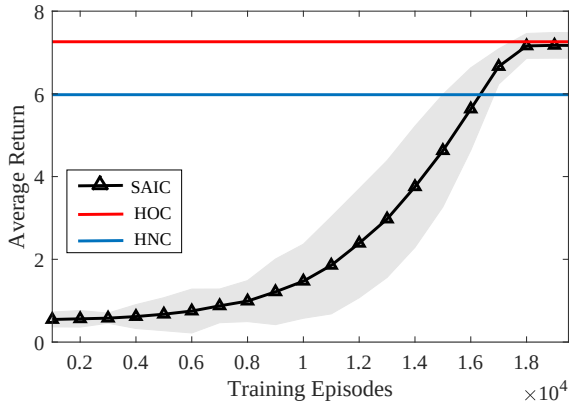

Figure 7. A comparison between SAIC, HOC and HNC within a three-agent system in terms of the system's average return with the bit-budget of $R = 1$ bit per time steps and number of training iterations/episodes $K = 20k$. The shaded area around SAIC's curve shows the standard deviation of SAIC in its performance.

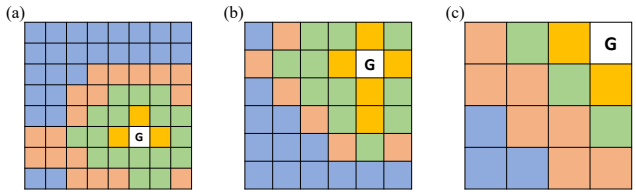

Figure 8. State aggregation for multi-agent communication in a two-agent rendezvous problem with grid-worlds of varied sizes and goal locations. The observation space is aggregated to four equivalence classes, $R = 2$ bits, and the number of training episodes has been $K = 1500k$, $K = 1000k$ and $K = 500k$ for figures (a) and (b) and (c) respectively. Locations with similar colours represent all the agents' observations which are grouped into the same equivalence class. The data compression ratio R_c has been seen to be 6:2, 5:2 and 4:2 in subplots a), b) and c) respectively. It is also observed that the observation clusters identified by SAIC have not been linearly separable under their original representation. In contrast, when clustered according to their values, observation points become linearly separable - see also Fig. 9 .

their values, as seen in Fig. 9, observation points become linearly separable. Fig. Fig. 9, allows us to see how precise the approximation of $V_{\pi^{m*},\pi^c}\big(\mathbf{o}_i(1), \mathbf{c}_{-i}(1)\big)$ by the value function $V^*\big(\mathsf{o}_i(t), \mathsf{c}_{-i}(t)\big)$ is - suggested by lemma 12. The figure illustrates the values for both $V_{\pi^{m*},\pi^c}\big(\mathbf{o}_i(1), \mathbf{c}_{-i}(1)\big)$ and $V^*\big(\mathsf{o}_i(t), \mathsf{o}_{-i}(t)\big)$, where $\mathsf{o}_i(t) = 21$ and $\mathsf{o}_{-i}(t)$ can take on possible values in Ω. For instance the values 7.2 mentioned on the right down corner of the grid demonstrates the value of $V^*\big(\mathsf{o}_i(t), \mathsf{o}_j(t)\big)$ when $\mathsf{o}_i(t) = 20$ and $\mathsf{o}_j(t) = 7$. This figure also allows finding the value of ϵ for all ϵ-cost-uniform groups.

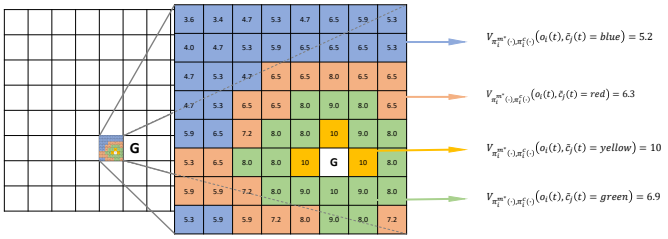

Figure 9. Left grid-world shows the observation space Ω, amongst which one particular observation is chosen $\mathsf{o}_i(t) = 20$. While agent i makes this observation, agent j can potentially be at any other 64 locations of the greed. The value function $V^*\big(\mathsf{o}_i(t) = 20, \mathsf{o}_j(t)\big)$ for all $\mathsf{o}_j(t) \in \Omega$ is depicted in the right grid-world, e.g. a number at location 22, shows the value function $V^*\big(\mathsf{o}_i(t) = 20, \mathsf{o}_j(t) = 22\big) = 10$. You can also see the values of $V_{\pi^{m*},\pi^c}\big(\mathbf{o}_i(t), \mathbf{c}_j(t)\big)$ for $\mathsf{o}_i(t) = 20$ and all possible $\mathsf{c}_j(t) \in \mathcal{C}$ with $R = 2$ bits.

We also investigate the impact of channel bit-budget R on the value of average return achieved by the LBIC, SAIC and CIC, in Fig. 10. In this figure, the normalized value of average return achieved for any scheme at any given R is shown. As per (22), the average return for the scheme of interest is computed by $\mathbb{E}_{p_{\pi^m,\pi^c}(\{\mathsf{tr}(t)\}_{t=1}^{t=M})}\{\mathbf{g}(1)\}$, where $\pi_i^m(\cdot)$ and $\pi_i^c(\cdot)$ are obtained by the scheme of interest after solving (10) with a given value of R. The average return is then normalized by dividing it to the average return $\mathbb{E}_{p_{\pi^*}(\{\mathsf{tr}(t)\}_{t=1}^{t=M})}\{\mathbf{g}(1)\}$ that is obtained by the optimal centralized policy $\pi^*(\cdot)$. The policy policy $\pi^*(\cdot)$ is the optimal solution to (3) under no communications constraint.

$$\frac{\mathbb{E}_{p_{\pi^m,\pi^c}(\{\mathsf{tr}(t)\}_{t=1}^{t=M})}\{\mathbf{g}(1)\}}{\mathbb{E}_{p_{\pi^*}(\{\mathsf{tr}(t)\}_{t=1}^{t=M})}\{\mathbf{g}(1)\}}. \tag{22}$$

Accordingly, when the normalized objective function of a particular scheme is seen to be close to the value 1, it implies that the scheme has been able to compress the observation information with almost zero loss with respect to the achieved objective function. On one hand, it is demonstrated that the SAIC achieves the optimal performance while running with 2 bits of inter-agent communications, while it takes the CIC at least $R = 4$ bits to get to achieve a sub-optimal value of the objective function. The LBIC, on the other hand, provides more than 10% performance gain in very low rates of communication $R \in \{1, 2, 3\}$ bits per time step, compared with CIC and 20% performance gain compared with SAIC at $R = 1$ bits per time step.

Fig. 11, studies the normalized objective functions attained by the LBIC, SAIC and CIC under different compression ratios R_c. A whopping 40% performance gain is acquired by the

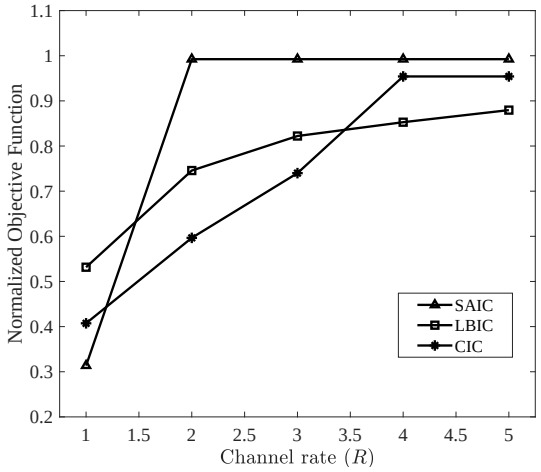

Figure 10. A performance comparison between several multi-agent communication and control schemes under different achievable bit rates. All experiments are performed where $N = 8$ and $\omega^T = 21$, similar to the grid-world of Fig. 8 -a. The number of training episodes/iterations for any scheme at any given channel bit-budget R has been $K = 200K$.

SAIC, in comparison to the CIC, at high compression ratio $R_c = 3 : 1$. This is equivalent to 66% of saving in the bit-budget with no performance drop with respect to the collaborative objective function. The SAIC, however, underperforms the LBIC and CIC at very high compression ratio of $R_c = 6 : 1$. This is due to the fact that the condition mentioned in remark 2 is not met at this high rate of compression. Moreover, the CIC scheme is seen not to achieve the optimal performance even at the compression rate of $R_c = 6 : 5$ which is due to the fact that by exceeding the compression ratio $R_c = 1 : 1$ each agent i may lose some information about the observation $\mathsf{o}_j(t)$ of the other agent which can be helpful in taking the optimal action decision.

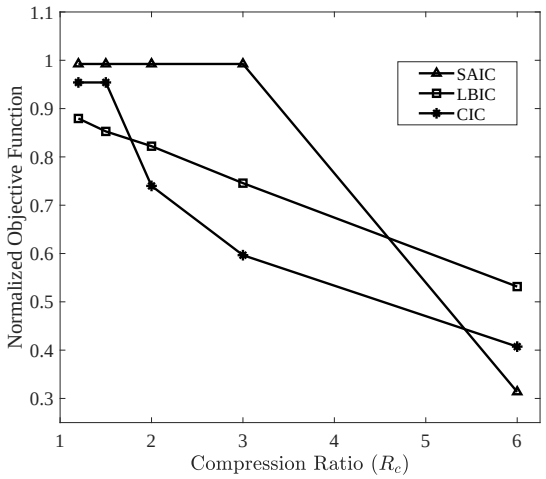

Figure 11. A performance comparison between several multi-agent communication and control schemes under different rates of data compression. All experiments are performed where $N = 8$ and $\omega^T = 21$. The number of training episodes/iterations for any scheme at any given bit-budget R has been $K = 200K$.

As demonstrated through a range of numerical experiments, the weakness of conventional schemes for compression of agents' observations is that they may lose/keep information regardless of how useful they can be towards achieving the optimal objective function. In contrast, the task-based compression schemes SAIC and LBIC, for communication bit-budgets (very) lower than the entropy of the observation process, manage to compress the observation information not to minimize the distortion but to maximize the achievable value of the objective function. Even though the numerical example provided in section IV, evaluates the performance of SAIC in a problem with a very low communication bit-budget, our theoretical results are applicable in scenarios with higher communication rates, as long as the processing unit that is deployed to solve the problem (3) is of sufficient computational resources to solve the problem in the desire time window.

VI. Conclusion

We have investigated the distributed joint design of communications and control for an MAS under bit-budgeted communications with the ultimate goal of maximizing the system's expected return. Since we consider a limited bit-budget for the multi-agent communication channels, task-based compression of agents' observations has been of the essence. Our proposed scheme, SAIC, which derives and solves the TODC problem can be differentiated from the conventional data quantization algorithms in the sense that it does not aim at achieving minimum possible distortion between the original signal and its reconstructed version - given a bit-budget for inter-agent communications. In contrast, SAIC aims at achieving the minimum possible distortion between the (learned) usefulness/value of the original observation signal and the learned usefulness/value of the the reconstructed observation signal - given a bit-budget for inter-agent communications. We have demonstrated the outstanding performance of SAIC compared with the conventional data compression algorithms, by up to a remarkable 40% improvement in the achieved objective function, when being imposed with tight constraints on the communication bit-budget.

To maximize the system's expected return, we could show analytically, how one can disentangle the TODC from the control problem - given the possibility of a centralized training phase. Our analytical studies confirm that despite the separation of the TODC and control problems, we can ensure very little compromise on the MAS's average return - compared with the jointly optimal control and quantization. Since the computational complexity of Q-learning in the centralized training phase is order of $|\Omega^n \times \mathcal{M}^n|$ time complexity [56], the addition of one single agent will multiply the complexity of the centralized training by $|\Omega \times \mathcal{M}|$. Thus, the complexity of the centralized training phase becomes a hurdle for the scalability of SAIC to a high number of agents. Accordingly, improving the scalability of the algorithm as well as extending the results for non-symmetric variable bit-budgets can be useful avenues to improve the applicability of the proposed schemes.

References

[1] A. Mostaani, T. X. Vu, S. K. Sharma, Q. Liao, and S. Chatzinotas, "Task-oriented communication system design in cyber-physical systems: A survey on theory and applications," *arXiv preprint arXiv:2102.07166*, 2021.

[2] D. Gunduz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. Kit Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *arXiv e-prints*, pp. arXiv–2207, 2022.

[3] E. C. Strinati and S. Barbarossa, "6g networks: Beyond shannon towards semantic and goal-oriented communications," *Computer Networks*, vol. 190, p. 107930, 2021.

[4] H. Witsenhausen, "Indirect rate distortion problems," *IEEE Transactions on Information Theory*, vol. 26, no. 5, pp. 518–521, 1980.

[5] P. Ioannou and J. Sun, "Theory and design of robust direct and indirect adaptive-control schemes," *International Journal of Control*, vol. 47, no. 3, pp. 775–813, 1988.

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

[7] H. Xie, Z. Qin, X. Tao, and K. B. Letaief, "Task-oriented multi-user semantic communications," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 9, pp. 2584–2597, 2022.

[8] N. Shlezinger and Y. C. Eldar, "Task-based quantization with application to mimo receivers," *arXiv preprint arXiv:2002.04290*, 2020.

[9] M. R. Palattella and N. Accettura, "Enabling internet of everything everywhere: Lpwan with satellite backhaul," in *2018 Global Information Infrastructure and Networking Symposium (GIIS)*. IEEE, 2018, pp. 1–5.

[10] L. Chaari, M. Fourati, and J. Rezgui, "Heterogeneous lorawan & leo satellites networks concepts, architectures and future directions," in *2019 Global Information Infrastructure and Networking Symposium (GIIS)*. IEEE, 2019, pp. 1–6.

[11] M. M. Azari, S. Solanki, S. Chatzinotas, O. Kodheli, H. Sallouha, A. Colpaert, J. F. M. Montoya, S. Pollin, A. Haqiqatnejad, A. Mostaani *et al.*, "Evolution of non-terrestrial networks from 5g to 6g: A survey," *IEEE Communications Surveys & Tutorials*, 2022.

[12] G. N. Nair and R. J. Evans, "Exponential stabilisability of finite-dimensional linear systems with limited data rates," *Automatica*, vol. 39, no. 4, pp. 585–593, 2003.

[13] ——, "Stabilizability of stochastic linear systems with finite feedback data rates," *SIAM Journal on Control and Optimization*, vol. 43, no. 2, pp. 413–436, 2004.

[14] M. Lauer and M. A. Riedmiller, "An algorithm for distributed reinforcement learning in cooperative multi-agent systems," in *Proc. Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., 2000.

[15] V. Kostina and B. Hassibi, "Rate-cost tradeoffs in control," *IEEE Transactions on Automatic Control*, vol. 64, no. 11, pp. 4525–4540, 2019.

[16] T.-Y. Tung, S. Kobus, J. R. Pujol, and D. Gunduz, "Effective communications: A joint learning and communication framework for multi-agent reinforcement learning over noisy channels," *arXiv preprint arXiv:2101.10369*, 2021.

[17] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 14–20, 1972.

[18] N. Shlezinger and Y. C. Eldar, "Deep task-based quantization," *Entropy*, vol. 23, no. 1, p. 104, 2021.

[19] 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.

[20] D. Lee, N. He, P. Kamalaruban, and V. Cevher, "Optimization for reinforcement learning: From a single agent to cooperative agents," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 123–135, 2020.

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

[22] F. Fischer, M. Rovatsos, and G. Weiss, "Hierarchical reinforcement learning in communication-mediated multiagent coordination," in *Proc. IEEE Joint Conference on Autonomous Agents and Multiagent Systems, 2004. AAMAS 2004.*, New York, Jul. 2004, pp. 1334–1335.

[23] T. Kasai, H. Tenmoto, and A. Kamiya, "Learning of communication codes in multi-agent reinforcement learning problem," in *Soft Computing in Industrial Applications, 2008. SMCia'08. IEEE Conf. on*. IEEE, 2008, pp. 1–6.

[24] F. Wu, S. Zilberstein, and X. Chen, "Online planning for multi-agent systems with bounded communication," *Artificial Intelligence*, vol. 175, no. 2, pp. 487–511, Feb. 2011.

[25] A. Amini, A. Asif, and A. Mohammadi, "Cease: A collaborative event-triggered average-consensus sampled-data framework with performance guarantees for multi-agent systems," *IEEE Transactions on Signal Processing*, vol. 66, no. 23, pp. 6096–6109, 2018.

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

[27] 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.

[28] A. Mostaani, T. X. Vu, S. Chatzinotas, and B. Ottersten, "State aggregation for multiagent communication over rate-limited channels," in *GLOBECOM 2020-2020 IEEE Global Communications Conference*. IEEE, 2020, pp. 1–7.

[29] D. Kim, S. Moon, D. Hostallero, W. J. Kang, T. Lee, K. Son, and Y. Yi, "Learning to schedule communication in multi-agent reinforcement learning," in *Intl. Conf. on Learning Representations*, 2019.

[30] R. Lowe, J. Foerster, Y.-L. Boureau, J. Pineau, and Y. Dauphin, "On the pitfalls of measuring emergent communication," in *Intl. Conf. on Autonomous Agents and MultiAgent Systems*, 2019.

[31] 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.

[32] 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.

[33] D. Abel, D. Hershkowitz, and M. Littman, "Near optimal behavior via approximate state abstraction," in *International Conference on Machine Learning*. PMLR, 2016, pp. 2915–2923.

[34] G. Rubino, "On weak lumpability in markov chains," *Journal of Applied Probability*, vol. 26, no. 3, pp. 446–457, 1989.

[35] D. Bertsekas, "Biased aggregation, rollout, and enhanced policy improvement for reinforcement learning," *arXiv preprint arXiv:1910.02426*, 2019.

[36] H. Zou, C. Zhang, S. Lasaulce, and et al, "Decision-oriented communications: Application to energy-efficient resource allocation," in *Intl. Conf. on Wireless Networks and Mobile Communications*. IEEE, 2018.

[37] H. Mao, Z. Zhang, Z. Xiao, Z. Gong, and Y. Ni, "Learning agent communication under limited bandwidth by message pruning," *arXiv preprint arXiv:1912.05304*, 2019.

[38] S. Sukhbaatar, R. Fergus *et al.*, "Learning multiagent communication with backpropagation," in *Proc. Advances in Neural Information Processing Systems*, Barcelona, 2016, pp. 2244–2252.

[39] P. A. Stavrou and M. Kountouris, "A rate distortion approach to goal-oriented communication," 2022.

[40] 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.

[41] G. E. Monahan, "State of the art—a survey of partially observable markov decision processes: theory, models, and algorithms," *Management science*, vol. 28, no. 1, pp. 1–16, 1982.

[42] P. Xuan, V. Lesser, and S. Zilberstein, "Communication decisions in multi-agent cooperation: Model and experiments," in *Proceedings of the Fifth International Conference on Autonomous Agents*, ser. AGENTS '01. New York, NY, USA: Association for Computing Machinery, 2001, p. 616–623. [Online]. Available: https://doi.org/10.1145/375735.376469

[43] F. A. Oliehoek, M. T. Spaan, N. Vlassis *et al.*, "DEC-PoMDPs with delayed communication," in *Proc. Multi-agent Sequential Decision-Making in Uncertain Domains*, Honolulu, Hawaii, May 2007.

[44] 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.

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

[46] Y. Rizk, M. Awad, and E. W. Tunstel, "Decision making in multiagent systems: A survey," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 3, pp. 514–529, 2018.

[47] C. Boutilier, "Multiagent systems: Challenges and opportunities for decision-theoretic planning," *AI magazine*, vol. 20, no. 4, pp. 35–35, 1999.

[48] T. Jaakkola, M. I. Jordan, and S. P. Singh, "Convergence of stochastic iterative dynamic programming algorithms," in *Advances in neural information processing systems*, 1994, pp. 703–710.

[49] F. Heylighen, "Stigmergy as a universal coordination mechanism i: Definition and components," *Cognitive Systems Research*, vol. 38, pp. 4–13, 2016.

[50] F. A. Oliehoek, C. Amato *et al.*, *A concise introduction to decentralized POMDPs*. Springer, 2016, vol. 1.

[51] S. Yüksel, "Jointly optimal lqg quantization and control policies for multi-dimensional systems," *IEEE Transactions on Automatic Control*, vol. 59, no. 6, pp. 1612–1617, 2013.
[52] S. Lloyd, "Least squares quantization in pcm," *IEEE transactions on information theory*, vol. 28, no. 2, pp. 129–137, 1982.
[53] Y. Linde, A. Buzo, and R. Gray, "An algorithm for vector quantizer design," *IEEE Transactions on communications*, vol. 28, no. 1, pp. 84–95, 1980.
[54] 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.
[55] C. Amato, J. S. Dibangoye, and S. Zilberstein, "Incremental policy generation for finite-horizon dec-pomdps," in *Nineteenth International Conference on Automated Planning and Scheduling*, 2009.
[56] M. G. Azar, R. Munos, M. Ghavamzadaeh, and H. J. Kappen, "Speedy q-learning," 2011.

APPENDIX A
PROOF OF THEOREM 1

To prove this theorem we first introduce a definition in subsection A-A, together with two lemmas and their proofs in subsections A-B and A-C. Lastly, we complete the proof of Theorem 1, in subsection A-D leveraging the above-mentioned.

A. Task-based information compression problem: a definition

Definition 10. *[Task-based information compression (TBIC) problem] Let the higher order function Π^{m^*} be a map from the vector space $\mathcal{K}^c$ of all possible joint communication policies $\pi^c = \langle \pi_1^c(\cdot), ..., \pi_n^c(\cdot)\rangle$ to the vector space $\mathcal{K}^m$ of optimal corresponding joint control policies $\pi^m = \langle \pi_1^{m^*}(\cdot), ..., \pi_n^{m^*}(\cdot)\rangle$. Upon the availability of Π^{m^*}, by plugging it into the problem (10), we will have a new problem*

$$\begin{aligned} &\underset{\pi_i^c}{max} && \mathbb{E}_{p_{\Pi^{m^*},\pi^c}\left(\{\mathsf{tr}(t)\}_{t=1}^{t=M}\right)}\Big\{\mathbf{g}(1)\Big\}, \quad i \in \mathcal{N} \\ &s.t. && log_2|\mathcal{C}| \leq R, \end{aligned} \tag{23}$$

where we maximize the system's return only with respect to the joint communication policies π^c. The joint optimal control policies $\langle \pi_1^{m^*}(\cdot), ..., \pi_n^{m^*}(\cdot)\rangle$ are automatically computed by the mapping $\Pi^{m^*}\big(\pi_1^c(\cdot), ..., \pi_n^c(\cdot)\big)$. The problem is called here as the TBIC problem.

B. Reformulating the objective function: a lemma

Lemma 11. *The objective function of the decentralized problem (10) can be expressed as*

$$\begin{aligned} &\mathbb{E}_{p_{\pi^m,\pi^c}(\{\mathsf{tr}(t)\}_{t=t'}^{t=M})}\big\{\mathbf{g}(t')\big\} = \\ &\mathbb{E}_{p_{\pi^m,\pi^c}(h_i(\mathbf{s}(t')))}\Big\{\mathbb{E}_{p_{\pi^m,\pi^c}(\{\mathsf{tr}(t)\}_{t=2}^{t=M}|h_i(\mathbf{s}(t')))}\big\{\mathbf{g}(t')|h_i(\mathbf{s}(t'))\big\}\Big\} = \\ &\mathbb{E}_{p_{\pi^m,\pi^c}(h_i(\mathbf{s}(t')))}\Big\{V_{\pi^m,\pi^c}\big(h_i(\mathbf{s}(t'))\big)\Big\}, \end{aligned} \tag{24}$$

for all $i \in \mathcal{N}$, where $V_{\pi^m,\pi^c}\big(h_i(\mathbf{s}(t'))\big)$ is the solution to the Bellman equation corresponding to the joint control and communication policies π^m, π^c.

Proof. Considering the definition of the value function, given in (25), the proof is immediately concluded when applying Adam's law on the expectation of the value function

$$V_{\pi^m,\pi^c}\big(h_i\big(\mathsf{s}(t')\big)\big) = \mathbb{E}_{p_{\pi^m,\pi^c}(\{\mathsf{tr}(t)\}_{t=t'+1}^{t=M})}\big\{\mathbf{g}(t')|h_i\big(\mathsf{s}(t')\big)\big\}. \tag{25}$$

■

C. Value of the perceived state of environment: a lemma

Lemma 12. *Using the knowledge of the solution $\pi^*(\cdot)$ to the centralized problem, we can find the optimal value of a perceived state $V^*\big(h_i\big(\mathsf{s}(t)\big)\big)$ in terms of the value of the underlying state $V^*\big(\mathsf{s}(t)\big)$ by*

$$V^*\big(h_i\big(\mathsf{s}(t)\big)\big) = \sum_{\mathsf{o}_1(t)\in\Omega} ... \sum_{\mathsf{o}_n(t)\in\Omega} V^*\big(\mathsf{s}(t)\big)\ p\big(\mathsf{o}_{-i}(t)|\mathsf{c}_{-i}(t)\big). \tag{26}$$

Proof.

$$V^*\big(h_i\big(\mathsf{s}(t')\big)\big)) = \tag{27}$$
$$\mathbb{E}_{p\left(\{tr\}_{t'}^{M}|h_i\left(\mathsf{s}(t')\right)\right)}\Big\{\sum\nolimits_{t=t'}^{M}\gamma^{t-1}r\big(\mathbf{s}(t),\mathbf{m}(t)\big)\big|h_i\big(\mathsf{s}(t')\big)\Big\} =$$
$$\mathbb{E}_{p\left(\{\mathsf{tr}\}_{t'}^{M}|h_i\left(\mathsf{s}(t')\right)\right)}\Big\{\mathbf{g}(t^{'})|h_i\big(\mathsf{s}(t')\big)\Big\} = \tag{28}$$
$$\sum\nolimits_{\{\mathsf{tr}\}_{t'}^{M}}\mathbf{g}(t^{'})p\Big(\{\mathsf{tr}\}_{t'}^{M}|h_i\big(\mathsf{s}(t')\big)\Big),$$

where the conditional probability $p\Big(\{\mathsf{tr}\}_{t'}^{M}|h_i\big(\mathsf{s}(t')\big)\Big)$ can be extended following the law of total probabilities

$$V^*\big(h_i\big(\mathsf{s}(t')\big)\big) = \sum\nolimits_{\{\mathsf{tr}\}_{t'}^{M}}\mathbf{g}(t^{'})\Bigg[\sum_{\mathsf{o}_1(t)\in\Omega} ... \sum_{\mathsf{o}_n(t)\in\Omega}$$
$$p\Big(\{\mathsf{tr}\}_{t'}^{M}|\mathsf{o}_i(t^{'}),\mathsf{o}_{-i}(t^{'}),\mathsf{c}_{-i}(t^{'})\Big)p\big(\mathsf{o}_{-i}(t')|\mathsf{c}_{-i}(t')\big)\Bigg], \tag{29}$$

where $\mathsf{o}_{-i}(t')$ is the observation vector of all agents $i \in \mathcal{N}_{-i}$. In eq. (29) $\mathsf{o}_i(t^{'}), \mathsf{o}_{-i}(t^{'})$ are sufficient statistics and can be replaced by $\mathsf{s}(t')$ and the second summation can be shifted to have

$$V^*\big(h_i\big(\mathsf{s}(t')\big)\big) = \sum_{\mathsf{o}_1(t)\in\Omega} ... \sum_{\mathsf{o}_n(t)\in\Omega}\sum_{\{\mathsf{tr}\}_{t'}^{M}}\mathbf{g}(t^{'})p\Big(\{\mathsf{tr}\}_{t'}^{M}|\mathsf{s}(t')\big)\Big)p\big(\mathsf{o}_{-i}(t)|\mathsf{c}_{-i}(t)\big), \tag{30}$$

where $\sum_{\{\mathsf{tr}\}_{t'}^{M}}\mathbf{g}(t^{'})p\Big(\{\mathsf{tr}\}_{t'}^{M}|\mathsf{s}(t')\big)\Big)$ can be replaced with $V^*\big(\mathsf{s}(t)\big)$, concluding the proof. ■

D. Proof of Theorem 1

Proof. Further to the result of lemma 11 and eq. (24), the original problem (10) can be expressed by

$$\begin{aligned} &\max_{\pi_i^m(\cdot),\pi_i^c(\cdot)} && \mathbb{E}_{p_{\pi^m,\pi^c}\left(h_i\left(\mathsf{s}(1)\right)\right)}\Big\{V_{\pi^m,\pi^c}\big(h_i\big(\mathsf{s}(1)\big)\big)\Big\}, \\ &\text{s.t.} && log_2|\mathcal{C}| \leq R, \end{aligned} \tag{31}$$

for $i \in \mathcal{N}$. Now by following definition 10 and plugging $\Pi^{m^*}(\cdot)$ into the problem (31) we obtain the TBIC problem

$$\begin{aligned} &\max_{\pi_i^c(\cdot)} && \mathbb{E}_{p_{\Pi^{m^*}(\pi^c),\pi^c}\left(h_i\left(\mathsf{s}(1)\right)\right)}\Big\{V_{\Pi^{m^*}(\pi^c),\pi^c}\big(h_i\big(\mathsf{s}(1)\big)\big)\Big\}, \\ &\text{s.t.} && log_2|\mathcal{C}| \leq R, \ \ i \in \mathcal{N}. \end{aligned} \tag{32}$$

We continue by following lemma 12, to be able to substitute $V_{\Pi^{m^*}(\pi^c),\pi^c}\big(h_i\big(\mathbf{s}(1)\big)\big)$ with its approximator $V^*\big(h_i\big(\mathbf{s}(1)\big)\big)$. This brings us to the approximated TBIC problem

$$\begin{aligned} \max_{\pi_i^c(\cdot)} \quad & \mathbb{E}_{p_{\pi^*,\pi^c}\left(h_i\left(\mathbf{s}(1)\right)\right)}\Big\{V^*\big(h_i\big(\mathbf{s}(1)\big)\big)\Big\} \;\; i \in \mathcal{N} \\ \text{s.t.} \quad & log_2|\mathcal{C}| \leq R. \end{aligned} \tag{33}$$

Note that the optimizers of the problem (33) and (34) are identical since the additional term $\mathbb{E}\big\{V^*\big(\mathbf{s}(t)\big)\big\}$ is independent from the communication policy $\pi_i^c(\cdot)$. Furthermore, the problem (34) is now expressed as a form of data quantization problem with mean absolute difference of the value functions $V^*\big(\mathbf{s}(t)\big)$ and $V^*\big(h_i\big(\mathbf{s}(1)\big)\big)$ as the measure of distortion. This interpretation of problem (34) can be better understood later by seeing the eq. (35).

$$\begin{aligned} \min_{\pi_i^c(\cdot)} \quad & \mathbb{E}_{p_{\pi^m,\pi^c}\left(h_i\left(\mathbf{s}(1)\right)\right)}\Big\{V^*\big(\mathbf{s}(1)\big) - V^*\big(h_i\big(\mathbf{s}(1)\big)\big)\Big\} \\ \text{s.t.} \quad & log_2|\mathcal{C}| \leq R, \end{aligned} \tag{34}$$

and since $V^*\big(\mathbf{s}(1)\big)$ is always larger than $V^*\big(h_i\big(\mathbf{s}(1)\big)\big)$, the problem above can also be written as

$$\begin{aligned} \min_{\pi_i^c(\cdot)} \quad & \mathbb{E}_{p_{\pi^m,\pi^c}\left(h_i\left(\mathbf{s}(1)\right)\right)}\Big\{\big|V^*\big(\mathbf{s}(1)\big) - V^*\big(h_i\big(\mathbf{s}(1)\big)\big)\big|\Big\} \\ \text{s.t.} \quad & log_2|\mathcal{C}| \leq R, \end{aligned} \tag{35}$$

concluding the proof of Theorem 1. ■

Appendix B
Proof of Lemma 2

Proof. The term $\mathbb{E}_{p_{\pi^m,\pi^c}\left(h_i\left(\mathbf{s}(1)\right)\right)}\Big\{V^*\big(\mathbf{s}(1)\big) - V^*\big(h_i\big(\mathbf{s}(1)\big)\big)\Big\}$ can be estimated by computing it over the empirical distribution of $\mathbf{s}(1)$. Note that the empirical joint distribution of $h_i\big(\mathbf{s}(1)\big)$ can be obtained by following the communication policy $\pi_i^c(\cdot)$ on the empirical distribution of $\mathbf{s}(1)$. Therefore, the problem (34) can be rewritten as

$$\begin{aligned} \min_{\pi_i^c(\cdot)} \quad & \sum_{\mathsf{o}_i(1)\in\Omega} ... \sum_{\mathsf{o}_n(1)\in\Omega} \Big|V^*\big(\mathbf{s}(t)\big) - V^*\big(h_i\big(\mathbf{s}(t)\big)\big)\Big|, \;\forall i \in \mathcal{N} \\ \text{s.t.} \quad & log_2|\mathcal{C}| \leq R. \end{aligned} \tag{36}$$

Quantization levels are disjoint sets $\mathcal{P}_{i,k} \subset \Omega$, where their union $\cup_{k=1}^{2^R}\mathcal{P}_{i,k}$ will cover the entire Ω. Each quantization level is represented by only one communication message $\mathsf{c}_j(t) = \mathsf{c}_k \in \mathcal{C}$. Further to lemma 12, the value of $V^*\big(h_i\big(\mathbf{s}(t)\big)\big)$ can be computed by empirical mean (26).

The quantization problem (36) becomes a k-median clustering problem

$$\min_{\mathcal{P}_i} \quad \sum_{\substack{\mathsf{o}_j(t)\in\Omega \\ j\in\mathcal{N}_{-i}}} \sum_{k=1}^{2^R} \sum_{\mathsf{o}_i(t)\in\mathcal{P}_{i,k}} \Big|V^*\big(\mathsf{o}_i(t),\mathsf{o}_j(t)\big) - \mu_k\Big|, \tag{37}$$

where $\mathcal{P}_i = \{\mathcal{P}_{i,1},...,\mathcal{P}_{i,2^R}\}$ is a partition of Ω, and the first summation $\sum_{\substack{\mathsf{o}_j(t)\in\Omega \\ j\in\mathcal{N}_{-i}}}$ is a concatenation of $n-1$ summations each one acting over $\mathsf{o}_j(t) \in \Omega$ where $j \in \mathcal{N}_{-i}$.

By taking the mean of $V^*\big(\mathsf{s}(t)\big)$ over the empirical distribution of $\mathsf{o}_j(t)$, $\forall j \in \mathcal{N}_i$, we can also marginalize out $\mathsf{o}_j(t)$, $\forall j \in \mathcal{N}_i$. Again, it does not change the solution of the problem and we will have

$$\min_{\mathcal{P}_i} \quad \sum\nolimits_{k=1}^{2^R} \sum\nolimits_{\mathsf{o}_i(t)\in\mathcal{P}_{i,k}} \Big|V^*\big(\mathsf{o}_i(t)\big) - \mu_k^{'}\Big|, \tag{38}$$

in which $\mu_k^{'} = \sum_{\mathsf{o}_j(t)\in\mathcal{P}_{i,k}} \mu_k$ will approximate $V^*\big(\mathsf{c}_i(t)\big)$. ■

To gain more insight about the meaning of this task-based information compression, it is useful to take a look at the conventional quantization problem which is adapted to our problem setting in eq. (39), where $\mathbf{c}_j = \pi_j^c\big(\mathbf{o}_j(1)\big)$. In fact, the compression scheme applied in the CIC, explained in subsection (V-B), is obtained by solving the following problem

$$\min_{\pi_i^c(\cdot)} \sum\nolimits_{\mathsf{o}_i(1)\in\Omega} \Big|\mathsf{o}_i(t) - \mathsf{c}_i(t)\Big|^2, \;\; \text{s.t. } log_2|\mathcal{C}| \leq R, \tag{39}$$

which can be solved optimally by the Lloyd's algorithm [52].

Appendix C
Proof of Lemma 4

Proof. Further to the law of iterated expectations, $V^*\big(\mathsf{o}_i(t')\big)$ can be expressed as

$$\begin{aligned} V^*\big(\mathsf{o}_i(t')\big) = \mathbb{E}_{p(\mathbf{o}_{-i}(t'))}\Bigg\{\mathbb{E}_{p_{\pi^*}(\{\mathsf{tr}(t)\}_{t=t'+1}^{t=M}|\mathsf{o}_i(t'),\mathbf{o}_{-i}(t'))}\Bigg\{ \\ \mathbf{g}(t')|\mathbf{o}_i(t') = \mathsf{o}_i(t'),\mathbf{o}_{-i}(t')\Bigg\}\Bigg\} = \\ \sum_{\mathsf{o}_{-i}(t')\in\Omega^{n-1}} p(\mathbf{o}_{-i}(t) = \mathsf{o}_{-i}(t'))\mathbb{E}_{\pi^*}\Big\{\mathbf{g}(t')|\mathsf{o}_i(t'),\mathsf{o}_{-i}(t')\Big\} \end{aligned} \tag{40}$$

where the expectation of the last term is the optimal value of the state $\mathsf{s}(t') = \langle \mathsf{o}_i(t'),\mathsf{o}_{-i}(t')\rangle$ of the underlying MDP

$$V^*\big(\mathsf{s}(t')\big) = \mathbb{E}_{\pi^*}\Big\{\mathbf{g}(t')|\mathsf{o}_i(t'),\mathsf{o}_{-i}(t')\Big\}. \tag{41}$$

Following Bellman optimality equation $V^*\big(\mathsf{s}(t')\big)$ can be obtained by centralized Q-learning following

$$\begin{aligned} V^*\big(\mathsf{s}(t')\big) &= \max_{\mathsf{m}\in\mathcal{M}^n} Q^*\big(\mathsf{s}(t'),\mathsf{m}(t')\big) \\ &= \mathbb{E}_{p_{\pi^*}(\{\mathsf{tr}(t)\}_{t=t'+1}^{t=M}|\mathsf{o}_i(t'),\mathsf{o}_{-i}(t'))}\Big\{\mathbf{g}(t')|\mathsf{o}_i(t'),\mathsf{o}_{-i}(t')\Big\}. \end{aligned} \tag{42}$$

Using (40) and (42) we can simply compute $V^*\big(\mathsf{o}_i(t')\big)$ by

$$V^*\big(\mathsf{o}_i(t)\big) = \sum_{\mathsf{o}_{-i}(t)\in\Omega^{n-1}} \max_{\mathsf{m}} \; Q^*\big(\mathsf{s}(t),\mathsf{m}(t)\big)p\big(\mathbf{o}_{-i}(t) = \mathsf{o}_{-i}(t)\big). \tag{43}$$

■

Appendix D
Proof of Theorem 8

Proof. Without loss of generality, we have written the proof of this theorem for a two agent scenario to improve the readability. Given the proof for the two-agent system, the extension to a multi-agent system is straightforward. According to the [33](Lemma 1), optimal state values of the aggregated MDPs (the environment as is seen by one agent during the decentralized training phase of SAIC) are in a

small neighbourhood of the optimal values corresponding to the optimal solution to the original underlying MDP:

$$\forall \mathsf{o}_j \in \Omega \text{ and } \text{ and } \forall i \in \{1,2\}, j \neq i : \\ |V^*(\mathsf{o}_i, \mathsf{o}_j) - V_i^m(\mathsf{o}_i, c_j^{(k)})| < \frac{2\epsilon}{(1-\gamma)^2}, \tag{44}$$

where $V_i^m(\cdot)$ is the value function corresponding to $\pi_i^{m,SAIC}(\cdot)$. The communication signal $c_j^{(k)} \in \mathcal{C}$ is agent j's communicated message and at the same time is the k-th element of the communication space $\mathcal{C} = \{c^{(1)}, c^{(2)}, ..., c^{|\mathcal{C}|}\}$ i.e., $c_j^{(k)} = c^{(k)}$. Following the eq. (24), one can write the expected return of the system under centralized scheme as :

$$\mathbb{E}_{p_{\pi^*}(\{\mathsf{tr}(t)\}_{t=t_0}^{t=M})}\{\mathbf{g}(t_0)\} = \mathbb{E}\Big\{V^*\big(\mathbf{o}_i(t_0), \mathbf{o}_j(t_0)\big)\Big\} = \\ \sum_{\mathbf{o}_j \in \Omega} \sum_{\mathbf{o}_i \in \Omega} V^*\big(\mathbf{o}_i(t_0), \mathbf{o}_j(t_0)\big) p_{\mathbf{o}_i,\mathbf{o}_j}(\mathsf{o}_i(t_0), \mathsf{o}_j(t_0)), \tag{45}$$

where the second expectation is taken over the joint probability distribution $p_{\pi^*}(\mathsf{o}_i(t_0), \mathsf{o}_j(t_0))$ of $\mathbf{o}_i$ and $\mathbf{o}_j$ when following the action policy $\pi^*(\cdot)$. This equation can be extended for multi-agent case only by taking a summation over each agent's observation space on the left-hand side. Similarly, following the eq. (24), one can write the expected return of the system that is run by SAIC as:

$$\mathbb{E}_{p_{\pi^m,\pi^c}(\{\mathsf{tr}(t)\}_{t=t_0}^{t=M})}\{\mathbf{g}(t_0)\} = \mathbb{E}\Big\{V^m\big(\mathbf{o}_i(t_0), \mathbf{c}_j^{(k)}(t_0)\big)\Big\} = \\ \sum_{k=1}^{|\mathcal{C}|} \sum_{\mathbf{o}_i \in \Omega} V^m\big(\mathbf{o}_i(t_0), \mathbf{c}_j^{(k)}(t_0)\big) p_{\mathbf{o}_i,\mathbf{c}_j}(\mathsf{o}_i(t_0), \mathsf{c}_j^{(k)}(t_0)). \tag{46}$$

We can rewrite the joint probability $p_{\mathbf{o}_i,\mathbf{c}_j}(\mathsf{o}_i(t_0), \mathsf{c}_j^{(k)}(t_0))$ as

$$p_{\mathbf{o}_i,\mathbf{c}_j}(\mathsf{o}_i(t_0), \mathsf{c}_j^{(k)}(t_0)) = \sum_{\mathsf{o}_j(t_0) \in \mathcal{P}_{i,k}} p_{\mathbf{o}_i,\mathbf{o}_j}(\mathsf{o}_i(t_0), \mathsf{o}_j(t_0)), \tag{47}$$

where the subset $\mathcal{P}_{i,k} \subset \Omega$ stands for the set of all observation realizations o_j that are represented by $\mathsf{c}_j^{(k)}(t_0)$ according to the policy $\pi_i^{c,SAIC}(\cdot)$. Given eq. (47), one can express eq. (46) - the expected return of the MAS under SAIC - also as

$$\mathbb{E}_{p_{\pi^m,\pi^c}(\{\mathsf{tr}(t)\}_{t=t_0}^{t=M})}\{\mathbf{g}(t_0)\} = \mathbb{E}\Big\{V^m\big(\mathbf{o}_i(t_0), \mathbf{c}_j^{(k)}(t_0)\big)\Big\} = \\ \sum_{k=1}^{|\mathcal{C}|} \sum_{\mathsf{o}_j(t_0) \in \mathcal{P}_{i,k}} \sum_{\mathbf{o}_i \in \Omega} V^m\big(\mathbf{o}_i(t_0), \mathbf{c}_j^{(k)}(t_0)\big) p_{\mathbf{o}_i,\mathbf{o}_j}(\mathsf{o}_i(t_0), \mathsf{o}_j(t_0)). \tag{48}$$

In order for eq. (45) to have the arrangement of its summations similar to eq. (48), it is sufficient to break its left-hand summation to two parts

$$\mathbb{E}_{p_{\pi^*}(\{\mathsf{tr}(t)\}_{t=t_0}^{t=M})}\{\mathbf{g}(t_0)\} = \mathbb{E}\Big\{V^*\big(\mathbf{o}_i(t_0), \mathbf{o}_j(t_0)\big)\Big\} = \\ \sum_{k=1}^{|\mathcal{C}|} \sum_{\mathsf{o}_j(t_0) \in \mathcal{P}_{i,k}} \sum_{\mathbf{o}_i \in \Omega} V^*\big(\mathbf{o}_i(t_0), \mathbf{o}_j(t_0)\big) p_{\mathbf{o}_i,\mathbf{o}_j}(\mathsf{o}_i(t_0), \mathsf{o}_j(t_0)), \tag{49}$$

Further to equations (49)-(48), the difference between the achievable expected return of the centralized scheme and SAIC can be explained by

$$\mathbb{E}_{p_{\pi^*}(\{\mathsf{tr}(t)\}_{t=t_0}^{t=M})}\{\mathbf{g}(t_0)\} - \mathbb{E}_{p_{\pi_i^m,\pi_i^c}(\{\mathsf{tr}(t)\}_{t=t_0}^{t=M})}\{\mathbf{g}(t_0)\} = \\ \sum_{k=1}^{|\mathcal{C}|} \sum_{\mathsf{o}_j(t_0) \in \mathcal{P}_{i,k}} \sum_{\mathbf{o}_i \in \Omega} V^*\big(\mathbf{o}_i(t_0), \mathbf{o}_j(t_0)\big) p_{\mathbf{o}_i,\mathbf{o}_j}(\mathsf{o}_i(t_0), \mathsf{o}_j(t_0)) - \\ \sum_{k=1}^{|\mathcal{C}|} \sum_{\mathsf{o}_j(t_0) \in \mathcal{P}_{i,k}} \sum_{\mathbf{o}_i \in \Omega} V^m\big(\mathbf{o}_i(t_0), \mathbf{c}_j^{(k)}(t_0)\big) p_{\mathbf{o}_i,\mathbf{o}_j}(\mathsf{o}_i(t_0), \mathsf{o}_j(t_0)). \tag{50}$$

We now proceed by factorizing the joint probability $p_{\mathbf{o}_i,\mathbf{o}_j}(\mathsf{o}_i(t_0), \mathsf{o}_j(t_0))$ which yields

$$\mathbb{E}_{p_{\pi^*}(\{\mathsf{tr}(t)\}_{t=t_0}^{t=M})}\{\mathbf{g}(t_0)\} - \mathbb{E}_{p_{\pi_i^m,\pi_i^c}(\{\mathsf{tr}(t)\}_{t=t_0}^{t=M})}\{\mathbf{g}(t_0)\} = \\ \sum_{k=1}^{|\mathcal{C}|} \sum_{\mathsf{o}_j(t_0) \in \mathcal{P}_{i,k}} \sum_{\mathbf{o}_i \in \Omega} p_{\mathbf{o}_i,\mathbf{o}_j}(\mathsf{o}_i(t_0), \mathsf{o}_j(t_0)) \big[V^*\big(\mathbf{o}_i(t_0), \mathbf{o}_j(t_0)\big) \\ - V^m\big(\mathbf{o}_i(t_0), \mathbf{c}_j^{(k)}(t_0)\big)\big] \tag{51}$$

Since $\big[V^*\big(\mathbf{o}_i(t_0), \mathbf{o}_j(t_0)\big) - V^m\big(\mathbf{o}_i(t_0), \mathbf{c}_j(t_0)\big)\big]$ is upper-bounded by a constant term $\frac{2\epsilon}{(1-\gamma)^2}$, its weighted sum is also upper bounded by the same term $\frac{2\epsilon}{(1-\gamma)^2}$. Thus we conclude the proof of Theorem 8. We are unsure if the suggested bound is tight. The results obtained in the performance evaluation indicates a large difference between the bound offered above and the performance bound between SAIC and the optimal centralized control. ■