

# Task-Effective Compression of Observations for the Centralized Control of a Multi-agent System Over Bit-Budgeted Channels

Arsham Mostaani, *Student Member, IEEE*, Thang X. Vu, *Senior Member, IEEE*,  
Symeon Chatzinotas, *Fellow, IEEE*, and Björn Ottersten, *Fellow, IEEE*

**Abstract**—We consider a task-effective quantization problem that arises when multiple agents are controlled via a centralized controller (CC). While agents have to communicate their observations to the CC for decision-making, the bit-budgeted communications of agent-CC links may limit the task-effectiveness of the system which is measured by the system's average sum of stage costs/rewards. As a result, each agent should compress/quantize its observation such that the average sum of stage costs/rewards of the control task is minimally impacted. We address the problem of maximizing the average sum of stage rewards by proposing two different Action-Based State Aggregation (ABSA) algorithms that carry out the indirect and joint design of control and communication policies in the multi-agent system. While the applicability of ABSA-1 is limited to single-agent systems, it provides an analytical framework that acts as a stepping stone to the design of ABSA-2. ABSA-2 carries out the joint design of control and communication for a multi-agent system. We evaluate the algorithms - with average return as the performance metric - using numerical experiments performed to solve a multi-agent geometric consensus problem. The numerical results are concluded by introducing a new metric that measures the effectiveness of communications in a multi-agent system.

**Index Terms**—Semantic communications, task-effective data compression, goal-oriented communications, communications for machine learning, multi-agent systems, reinforcement learning.

## I. INTRODUCTION

As 5G is rolling out, a wave of new applications such as the internet of things (IoT), industrial internet of things (IIoT) and autonomous vehicles is emerging. It is projected that by 2030, approximately 30 billion IoT devices will be connected [1]. With the proliferation of non-human types of connected devices, the focus of the communications design is shifting from traditional performance metrics, e.g., bit error rate and latency of communications to the semantic and task-oriented performance metrics such as meaning/semantic error rate [2], [3] and the timeliness of information [4]. To evaluate how efficiently the network resources are being utilized, one could traditionally measure the sum rate of a network whereas in the era of the cyber-physical systems, given the resource constraints of the network, we want to understand how effectively one can conduct a (number of) task(s) in the desired way [5], [6]. We are witnessing a paradigm shift in communication systems where the targeted performance metrics of the traditional systems are no longer valid. This imposes new grand challenges in designing the

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 European Research Council (ERC) via the project AGNOSTIC, ref. EC/H2020/ERCAdG/742648, and the Luxembourg National Fund (FNR) via project RUTINE, ref. FNR/C22/IS/17220888.

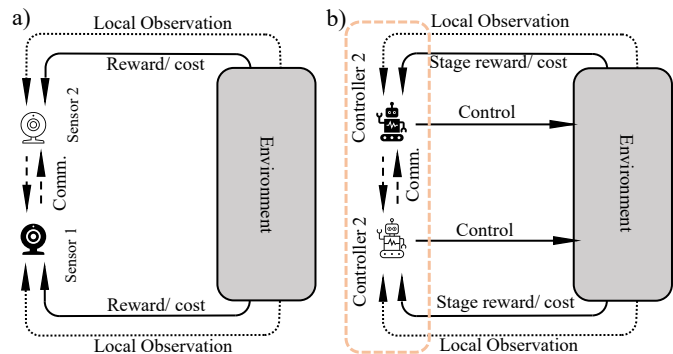


Figure 1. Task-effective communications for a) an estimation vs. b) a control task - the orange dashed box is detailed in Fig. 2 and Fig. 3.

communications towards the eventual task-effectiveness [6]. This line of research is also driven partly due to the success of new machine learning technologies/ algorithms under the title of "emergent communications" in multi-agent systems [7]. Transfer of these new technologies/ideas to communication engineering is anticipated to have a disruptive effect in multiple domains of the design of communication systems.

According to Shannon and Weaver, communication problems can be divided into three levels [8]: (i) technical problem: given channel and network constraints, how accurately can the communication symbols/bits be transmitted? (ii) semantic problem: given channel and network constraints, how accurately the communication symbols can deliver the desired meaning? (iii) effectiveness problem: given channel and network constraints, how accurately the communication symbols can help to fulfil the desired task? While the traditional communication design addresses the technical problem, recently, the semantic problem [2], [3], [5], [9], [10] as well as the effectiveness problem [6], [11]–[18] have attracted extensive research interest.

In contrast to Shannon's technical-level communication framework, semantic communication can enhance performance by exploiting prior knowledge between source and destination [4], [19]. The semantic-based designs, however, are not necessarily task-effective [20]. One can design transmitters which compress the data with the least possible compromise on the semantic meaning being transmitted [2], [3] while the transmission can be task-unaware [21]. In contrast to semantic level and technical level communication design, the performance of a task-effective communication system is ultimately measured in terms of the average return/cost linked to the task [11]. In the (task) effectiveness problem, we are

not concerned only about the communication of meaning but also about how the message exchange is helping the receiving end to improve its performance in the expected cost/reward of an estimation task [4], [13], [14], [16], [22] or a control task [11], [12], [14], [15], [17], [18], [23], [24].

There are fundamental differences between the design of task-effective communications for an estimation vs. a control task - Fig. 1. (i) In the latter, each agent can produce a control signal that directly affects the next observations of the agent. Thus, in control tasks, the source of information - local observations of the agent - is often a stochastic process with memory - e.g. linear or Markov decision processes - [11], [17], [18]. In the estimation tasks, however, the source of information is often assumed to be an independent and identically distributed (i.i.d.) stochastic process [13], [16], [22]. (ii) In the control tasks, a control signal often has a long-lasting effect on the state of the system more than for a single stage/time step e.g., a control action can result in lower expected rewards in the short run but higher expected rewards in the long run. This makes the control tasks intrinsically sensitive to the time horizon for which the control policies are designed. Estimation tasks, specifically when the observation process is i.i.d., can be solved in a single stage/ time step - since there is no influence from the solution of one stage/ time step to another i.e., each time step can be solved separately [22], [25]. (iii) The cost function for estimation tasks is often in the form of a difference/distortion function while in the control tasks, it can take on many other forms.

In this paper, we focus on the effectiveness problem for the control tasks. In particular, we investigate the distributed communication design of a multiagent system (MAS) with the ultimate goal of maximizing the expected summation of per-stage rewards also known as the expected return. Multiple agents select control actions and communicate in the MAS to accomplish a collaborative task with the help of a central controller (CC) - i.e. the communication network topology of the MAS is a star topology with the hub node being the central controller and the peripheral nodes being the agents - Fig. 2. The considered system architecture can find applications in several domains such as the Internet of Things, emerging cyber-physical systems, real-time interactive systems, vehicle-to-infrastructure communication [26], collaborative perception [27] and vehicle-to-edge off-loading [28].

#### A. Related works: Task-effective communications for control tasks

Authors in [11], [12], [14], [17], [18], [23], [24] consider task-effective communication design under different settings. While [12], utilizes the task-effective communication design for the specific problem of the design of application-tailored protocols over perfect communication channels, the communication channel is considered to be imperfect in [11], [14], [17], [18], [23], [24]. Authors in [14] provide algorithmic contributions to the design of task-effective joint source-channel coding for single agent systems. Task-effective joint source and channel coding for MAS is targeted by [11],

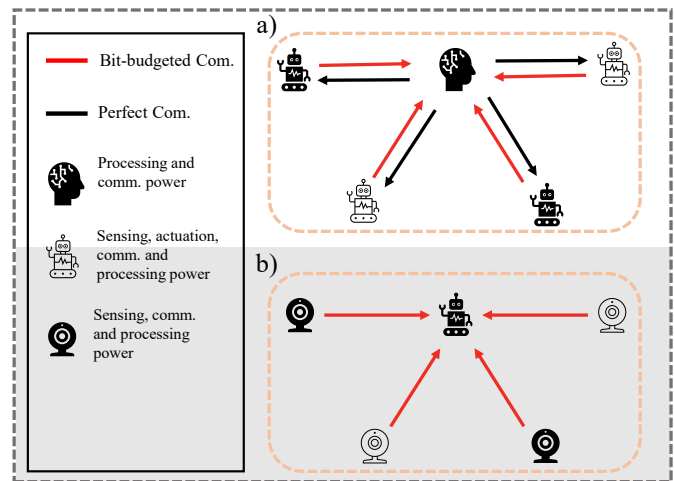


Figure 2. Communication topology and its applicable scenarios a) Centralized control of a MAS with collocated actuators and sensors, b) Distributed sensing with a single controller collocated with a single actuator. The orange dashed box is detailing the same box in Fig. 1 and Fig. 3 .

[14], [17], whereas [18], [23] are focused on task-effective data compression and quantization. Similar to the current paper, a star topology for the inter-agent communication is considered in [11], [12] whereas [12] assumes perfect communications between the hub node and the peripherals and [11] assumes imperfect communication channels at the down-link of the peripheral nodes. Despite the growing literature on task-effective communications for single and multi-agent systems, a part of which was surveyed above, a notable void persists in addressing the problem of the joint design of uplink quantization and control for a multi-agent system communicating over bit-budgeted channels of a star topology network. Diverse applications of this problem are mentioned in the last subsection. In contrast to all the above-mentioned work, this paper is - to the best of our knowledge - the first to study the star topology with the uplink (agent to hub) channel be imperfect (bit-budgeted) - Fig. 2. Equivalently, each agent observes the environment and communicates an abstract version of its local observation to the CC via imperfect (bit-budgeted) communication channels - red links in Fig. 2. Subsequently, CC produces control actions that are communicated to the agents via perfect communication channels - black links in Fig. 2. The control actions are selected by the CC such that they maximize the average return of the collaborative task, where the return is a performance metric linked to the accomplishment of the task.

#### B. Contributions

In our earlier work [18], we have developed a generic framework to solve task-oriented communication problems - for a multi-agent system (MAS) with full mesh connectivity. The current work can be considered as an adoption of that framework to a new problem setting for the design of task-effective communications where agents follow a star network topology for their connectivity. In this direction, the current work transcends the applicability of the proposed framework

beyond the specific problem that was solved in [18] and provides further insights into how the framework can be used in wider terms and under a wider range of settings. In particular, the contributions of this work are listed below.

- Firstly, we consider a novel problem setting in which a MAS is controlled via a central controller who has access to agents' local observations only through bit-budgeted distributed communications. This problem setting can be used in collaborative perception systems as well as vehicle-to-infrastructure communications, which cannot be addressed by the problem settings investigated in the prior similar art.
- Secondly, our analytical studies establish the relationship between the considered joint communication and control design problem and conventional data quantization problems. In particular, lemma 1 shows how the problem approached in this paper is a generalized version of the conventional data quantization. This formulation is useful as it helps to find an exact solution to the problem under stronger conditions via ABSA-1 and under milder conditions via ABSA-2.
- Moreover, our analytical studies help us to craft an indirect<sup>1</sup> task-effective data quantization algorithm - ABSA-2. Designing a task-effective data quantization for ABSA-2 can equivalently be translated as an indirect approach to feature selection for an arbitrary deep Q-network. Relying on the analysis carried out for ABSA-1, ABSA-2 designs distributed and bit-budgeted communications between the agents and CC. ABSA-2 is seen to approach optimal performance by increasing the memory of the CC. In fact, increasing the memory of CC leads to higher computational complexity. Therefore, ABSA-2 is said to strike a trade-off between computational complexity and task efficiency.
- Numerical experiments are carried out on a geometric consensus task to evaluate the performance of the proposed schemes in terms of the optimality of the MAS's expected return in the task. ABSA-1 and ABSA-2 are compared with several other benchmark schemes introduced by [18], in a multi-agent<sup>2</sup> scenario with local observability and bit-budgeted communications.
- Finally, we will introduce a new metric, called task-relevant information, for the measurement of effectiveness in task-oriented communication policies that - in comparison with the existing metrics such as positive listening and positive signalling - better explains the behaviour of a variety of task-effective communication schemes. The proposed metric is capable of measuring the effectiveness of a task-oriented communica-

<sup>1</sup>By an indirect algorithm here we mean an approach that is not dependent on our knowledge from a particular task. Indirect approaches are applicable to any/(wide range of) tasks. In contrast to indirect schemes, we have direct schemes that are specifically designed for a niche application [16]. As defined by [6]: "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".

<sup>2</sup>Due to the complexity related issues explained in section IV, the numerical results are limited to two-agent and three-agent scenarios.

tion/compression policy without the need of testing a jointly designed control policy and testing the jointly designed policies in the desired task.

### C. Technical approach

Our goal is to perform an efficient representation of the agents' local observations to ensure meeting the bit-budget of the communication links while minimizing the effect of quantization on the average return of the task. To achieve this, we first need to design task-effective data quantization policies for all agents. In task-effective data quantization, one needs to take into account the properties of the average return function and the optimal control policies associated with the task [15]. In addition to the design of the quantization policies for all agents, we also need the control policy of the CC to be capable of carrying out near-optimal decision-making despite its mere access to the quantized messages - resulting in a joint control and data compression problem. We formulate the joint control and data compression problem as a generalized form of data compression: task-oriented data compression (TODC). Following this novel problem formulation, we propose two indirect action-based state aggregation algorithms (ABSA): (i) ABSA-1 provides analytical proof for a task-effective quantization i.e., with optimal performance in terms of the expected return. In this direction, ABSA-1 relaxes the assumption of the lumpability of the underlying MDP, according to which [18][condition. 6], the performance guarantees of the proposed method were established. Since ABSA-1 is only applicable when the system is composed of one agent and the CC we also propose ABSA-2. Following the analytical results of ABSA-1, given the help of MAP estimation to relax the aforementioned limitation of ABSA-1, and benefit from a DQN controller at the CC; ABSA-2 will be introduced as a more general approach. (ii) ABSA-2 solves an approximated version of the TODC problem and carries out the quantization for any number of agents communicating with the CC. Thanks to a deep Q-network controller utilized at the CC, ABSA-2 can solve more complex problems where the controller benefits from larger memory. Thus, ABSA-2 allows trading complexity for communication efficiency and vice versa. Finally, we will evaluate the performance of the proposed schemes in the specific task: a geometric consensus problem under finite observability [29].

### D. Organization

The rest of this paper is organized as follows. Section II describes the MAS and states the joint control and communication problem. Section III proposes two action-based state aggregation algorithms. Section IV shows the performance of the proposed algorithms in a geometric consensus problem. Finally, Section V concludes the paper. For the reader's convenience, a summary of the notation that we follow in this paper is given in Table I. Bold font is used for matrices or scalars which are random and their realizations follow simple font.

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

Table I  
TABLE OF NOTATIONS

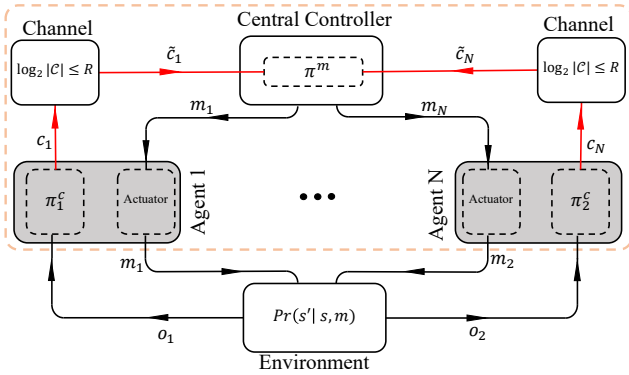


Figure 3. Illustration of the interactions of the CC and agents for the control of the environment. The red link shows the communication channels that are bit-budgeted - implying the local (and not global) observability of the CC. The orange dashed box is detailing the same box in Fig. 1 and Fig. 2 .

## II. SYSTEM MODEL AND PROBLEM STATEMENT

The problem setting we introduce here can be used to analyse both scenarios illustrated in Fig. 2. Nevertheless, to use our language consistently, we focus on the scenario (a) of that figure throughout the manuscript. In particular, when we use the term "agent" we refer to an object which certainly has all the following hardware capabilities: sensing, actuation, communication and data processing. A MAS, however, may not be comprised of mere agents, but of a combination of agents and perhaps other objects that has at least the hardware capabilities for communication and data processing power. The central controller here is supposed to have the hardware capability to process relatively larger data as well as the capability of communications. The interactions inside the MAS and outside the MAS with the environment are illustrated in Fig. 3.

### A. System model

We consider a MAS in which multiple agents  $i \in \mathcal{N} = \{1, 2, \dots, N\}$  collaboratively solve a task with the aid of a CC. Following a centralized action policy, CC provides the agents with their actions via a perfect communication channel while it receives the observations of agents through an imperfect

communication channel<sup>3</sup>. The considered setting is similar to conventional centralized control of MASs [18], [31], except for the fact that the communications from the agents to the CC are transmitted over a bit-budgeted communication channel. The agent-hub communications are considered to be instantaneous and synchronous [18]. This is in contrast with the delayed [17], [32] and sequential/iterative communication models [33]–[35]. We note that there is no direct inter-agent communication in the considered system - communications occur only between agents and the central controller. The system runs on discrete time steps  $t$ . The observation of each agent  $i$  at time step  $t$  is shown by  $\mathbf{o}_i(t) \in \Omega$  and the state  $\mathbf{s}(t) \in \mathcal{S}$  of the system is defined by the joint observations

$$\mathbf{s}(t) \triangleq \langle \mathbf{o}_1(t), \dots, \mathbf{o}_N(t) \rangle^4. \quad (1)$$

The control action of each agent  $i$  at time  $t$  is shown by  $\mathbf{m}_i(t) \in \mathcal{M}$ , and the action vector  $\mathbf{m}(t) \in \mathcal{M}^N$  of the system is defined by the joint actions  $\mathbf{m}(t) \triangleq \langle \mathbf{m}_1(t), \dots, \mathbf{m}_N(t) \rangle$ . The observation space  $\Omega$ , state-space  $\mathcal{S}$ , and action space  $\mathcal{M}$  are all discrete sets. The environment is governed by an underlying<sup>5</sup> Markov Decision Process that is described by the tuple  $M = \{\mathcal{S}, \mathcal{M}^N, r(\cdot), \gamma, T(\cdot)\}$ , where  $r(\cdot) : \mathcal{S} \times \mathcal{M}^N \rightarrow \mathbb{R}$  is the per-stage reward function and the scalar  $0 \leq \gamma \leq 1$  is the discount factor. The function  $T(\cdot) : \mathcal{S} \times \mathcal{M}^N \times \mathcal{S} \rightarrow [0, 1]$  is a conditional probability mass function (pmf) which represents state transitions such that  $T(\mathbf{s}(t+1), \mathbf{s}(t), \mathbf{m}(t)) = \Pr(\mathbf{s}(t+1)|\mathbf{s}(t), \mathbf{m}(t))$ . According to the per-stage reward signals, the system's return within the time horizon  $T'$  is denoted by

$$\mathbf{g}(t') = \sum_{t=t}^{T'} \gamma^{t-1} r(\mathbf{o}_1(t), \dots, \mathbf{o}_N(t), \mathbf{m}_1(t), \dots, \mathbf{m}_N(t)). \quad (2)$$

While the system state is jointly observable by the agents [36], each agent  $i$ 's observation  $\mathbf{o}_i(t)$  is local<sup>6</sup>. Once per time step, agent  $i \in \mathcal{N}$  is allowed to transmit its local observations through a communication message  $c_i(t)$  to the CC. The communications between agents and the central controller are done in a synchronous (not sequential) and simultaneous (not delayed) fashion [17]. Each agent  $i$  generates its communication message  $c_i(t)$  by following its communication policy  $\pi_i^c(\cdot) : \Omega \rightarrow \mathcal{C}$ . In parallel to all other agents, agent  $i$  follows the communication policy  $\pi_i^c(\cdot)$  to map its current observation  $\mathbf{o}_i(t)$  to the communication message  $c_i(t)$  which will be received by the central controller in the same time-

<sup>3</sup>In this work we follow a common assumption used in the networked control literature [30] according to which the bit-budget only limits the uplink communications of the agents and not their downlink. Accordingly, the agents select their control actions as dictated to them by the central controller.

<sup>4</sup>According to this definition, at any given time  $t$  the observations of any two agent  $i, j \in \mathcal{N}$  are linearly independent in the Euclidean space. The same conditions are true for the control actions of arbitrary agents.

<sup>5</sup>As defined in the literature [10], the underlying MDP<sup>\*</sup> 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  $\mathbf{s}(t) \triangleq \langle \mathbf{o}_1(t), \dots, \mathbf{o}_N(t) \rangle$  that has the same transition model  $T(\cdot)$  and reward model  $r(\cdot)$  as the environment experienced by our MAS.

<sup>6</sup>In our problem setting, each agent does not see the environment as an MDP due to their local observability. We only assume the presence of an underlying MDP for the environment, which is widely adopted in the literature for the reinforcement learning algorithm, e.g., [37] [38]. We have this assumption as our performance guarantees rely on the optimality of the solution provided for the control task, which is also assumed in [7], [10]. Let us recall that throughout all of our numerical studies, even the CC, given joint observations of all agents, cannot observe the true/nominal state of the environment.

step  $t$ . The code-book  $\mathcal{C}$  is a set composed of a finite number of communication code-words  $s, c, c', c'', \dots, c^{(|\mathcal{C}|-1)}$  - we use the same notation to refer to the different members of the action, observation and state spaces too. Agents' communication messages are sent over an error-free finite-rate bit pipe, with its rate constraint to be  $R \in \mathbb{R}$  (bits per channel use) or equivalently (bits per time step). As a result, the size of the quantization codebook should follow the inequality  $|\mathcal{C}| \leq 2^R$ . The CC exploits the received communication messages  $\mathbf{c}(t) \triangleq \langle \mathbf{c}_1(t), \dots, \mathbf{c}_N(t) \rangle$  within the last  $d$  number of time-steps to generate the action signal  $\mathbf{m}(t)$  following the control policy  $\pi^m(\cdot) : \mathcal{C}^{Nd} \rightarrow \mathcal{M}^N$ . Based on the above description, the environment from the point of view of the CC as well as from the agent's point of view is not necessarily an MDP - as none is capable of viewing the nominal state of the environment.

### B. Problem statement: Joint Control and Communication Design (JCCD) problem

Now we define the JCCD problem. Let  $M$  be the MDP governing the environment and the scalar  $R \in \mathbb{R}$  to be the bit-budget of the uplink of all agents. At any time step  $t'$ , we aim at selecting the tuple  $\pi = \langle \pi^m(\cdot), \pi^c \rangle$  with  $\pi^c \triangleq \langle \pi_1^c(\cdot), \dots, \pi_N^c(\cdot) \rangle$  to solve the following variational dynamic programming

$$\operatorname{argmax}_{\pi} \mathbb{E}_{\pi} \left\{ \mathbf{g}(t') \right\}; \quad \text{s.t. } |\mathcal{C}| \leq 2^R, \quad (3)$$

where the expectation is taken over the joint pmf of the system's trajectory  $\{\operatorname{tr}\}_{t'}^{T'} = \langle \mathbf{o}_1(t'), \dots, \mathbf{o}_N(t'), \mathbf{m}(t'), \dots, \mathbf{o}_1(T'), \dots, \mathbf{o}_N(T'), \mathbf{m}(T') \rangle$ , when the agents follow the policy tuple  $\pi$ . In the next section, similar to [18] we will disentangle the design of action and communication policies via action-based quantization of observations. In contrast to [18], here the communication network of the MAS is assumed to follow a star topology. The idea behind this disentanglement is to extract the features of the control design problem that can affect the communication design and to take them into account while designing the communications. Thus our communication design will be aware of the key features of the control task. We extract the key features of the control task using analytical techniques as well as reinforcement learning [17], [18]. In fact, the new communication problem called TODC, will no longer be similar to the conventional communication problems, as it is inspired by the JCCD problem.

In [18], [23], authors use the value of agents' observations for the given task as the key feature of the control task considered in the communication design. Accordingly, the idea was to cluster together the observation points that have similar values. In contrast to [18], [23], which considers the value of observations as an explicit key feature of the control task, here we consider the optimal control/action values assigned to each observation as the key feature. Accordingly, ABSA clusters the observation values together, whenever the observation points have similar optimal control/action values assigned to them. Action-based state aggregation has been already introduced in

the literature of reinforcement learning as a means for reducing the complexity of the reinforcement learning algorithms while maintaining the average return performance [39], [40].

### III. ACTION-BASED LOSSLESS COMPRESSION OF OBSERVATIONS

In this section, we will set yet another example - in addition to [18] - for the use of a generic framework to solve JCCD problem. In [18], a similar problem is solved for distributed control and quantization, wherein, the authors disentangle the design of task-oriented communication policies and action policies given the aid of a hypothetical functional  $\Pi^{m^*}$ . In particular, the functional  $\Pi^{m^*}$  is a map from the vector space  $\mathcal{K}^c$  of all possible communication policies  $\pi^c$  to the vector space  $\mathcal{K}^m$  of optimal corresponding control policy  $\pi^{m^*}(\cdot)$ . Upon the availability of the functional  $\Pi^{m^*}$ , wherever the function  $\pi^m$  appears in the JCCD problem, it can be replaced with  $\Pi^{m^*}(\pi^c)$  resulting in a novel problem in which only the communication policies  $\pi^c$  are to be designed. While in [18], authors use an approximation of  $\Pi^{m^*}(\pi^c)$  to obtain a task-oriented quantizer design problem, in the current work we derive an exact solution for a simplified version of (4) - where the number of agents communicating with the central controller is limited to one agent. To adapt ABSA to the generic setting of the problem (4), in ABSA-2, we will lift this limitation given the aid of an approximation technique.

The JCCD problem can already be formulated as a form of data-quantization problem. Lemma 1, identifies the quantization metric that we aim to optimize in this paper. It reformulates the JCCD problem as a novel generalized data quantization problem.

**Lemma 1.** *The JCCD problem (3) can also be expressed as a generalized data quantization problem as follows*

$$\operatorname{argmin}_{\pi} \mathbb{E}_{p(\mathbf{s}(t))} \left| V^{\pi^*}(\mathbf{s}(t)) - V^{\pi^m}(\mathbf{c}(t)) \right|, \quad \text{s.t. } |\mathcal{C}| \leq 2^R, \quad (4)$$

where the communication vector  $\mathbf{c}(t)$  generated by  $\pi^c$  is a quantized version of the system's state  $\mathbf{s}(t)$ .

*Proof.* Appendix A. ■

In contrast to the classic data-quantization problems, here the distortion metric, measures the difference between two different functions of the original signal and its quantized version - namely  $V^{\pi^*}(\cdot)$  and  $V^{\pi^m}(\cdot)$  - thus the distortion measure that we aim to optimize by solving (4) is not conventional. In fact, the variational minimization problem is solved over the vector space of joint quantization policies  $\pi^c$  and action policy  $\pi^m$  functions.

#### A. ABSA-1 Algorithm

The applicability of the proposed ABSA-1 is limited to two mathematically equivalent scenarios: (i) we have a single agent communicating to the CC - consider the Fig. 2-a, with only one agent connected to the CC - or (ii) that the agents communicate with the CC through a relay. In the latter scenario, the relay has full access to the agents' communication observation, i.e.,  $\mathbf{o}_i, \forall i \in \mathcal{N}$ , while the relay to CC channel is bit-budgeted.

This limited scenario is useful for us to facilitate our analytical studies on the problem (4), allowing us to establish theoretical proof for the losslessness of compression in ABSA-1 as well as its optimal average return performance. These statements will be confirmed by Lemma 2 - the results of which will also be useful to design ABSA-2. The central idea of ABSA-1 is to represent any two states  $\mathbf{s}^{(i)}, \mathbf{s}^{(j)}$  using the same communication message  $\mathbf{c}$  iff  $\pi^*(\mathbf{s}^{(i)}) = \pi^*(\mathbf{s}^{(j)})$ , where  $\pi^*(\cdot) : \mathcal{S} \rightarrow \mathcal{M}^N$  is the optimal control policy of the agents, given the access of observations from all agents. Thus, ABSA-1 and ABSA-2 solve the JCCD problem at three different phases: (i) solving the centralized control problem under perfect communications via reinforcement learning i.e., Q-learning, to find  $\pi^*(\cdot)$ <sup>7</sup>, (ii) solving the task-oriented data quantization problem to find  $\pi^c$  via a form of data clustering, (iii) finding the  $\pi^m$  corresponding to  $\pi^c$ .

In order to explain ABSA-1, we introduce the problem of task-oriented data compression with centralized control. TBIC is derived using similar techniques in [18] but for a different setting i.e., the communication network of MAS has a star topology. The TBIC problem is no longer a joint control and communication problem but is a quantization design problem in which the features of the control problem are taken into account. To arrive to TODC problem from the JCCD problem, we use the functional  $\Pi^{m^*}$  to replace  $\pi^m(\cdot)$  with  $\Pi^{m^*}(\pi^c)$ . Upon the availability of  $\Pi^{m^*}$ , by plugging it into the JCCD problem (3), we will have a new problem

$$\operatorname{argmin}_{\pi^c} \mathbb{E}_{p(\mathbf{s}(t))} \left| V^{\pi^*}(\mathbf{s}(t)) - V^{\Pi^{m^*}(\pi^c)}(\mathbf{c}(t)) \right|, \quad \text{s.t. } |\mathcal{C}| \leq 2^R, \quad (5)$$

where we maximize the system's return with respect to only the communication policies  $\pi^c(\cdot)$  of the local relay. The optimal control policy  $\pi^{m^*}(\cdot)$  of the CC is automatically computed by the mapping  $\Pi^{m^*}(\pi^c(\cdot))$ . The problem is called here as the TODC problem. Upon the availability of  $\Pi^{m^*}$ , the JCCD problem (3) can be reduced to (5). Definition 1 is provided to formalize a precise approach to solve (5) via obtaining the communication policy of the relay  $\pi^c(\cdot)$  as well as the corresponding  $\Pi^{m^*}$ , to solve (3).

**Definition 1. Quantization and control policies in ABSA-1:**

The communication policy  $\pi^{c, ABSA-1}(\cdot)$  designed by ABSA-1 will be obtained by solving the following  $k$ -median clustering problem

$$\min_{\mathcal{P}} \sum_{i=1}^{|\mathcal{C}|} \sum_{\mathbf{s}(t) \in \mathcal{P}_i} \left| \pi^*(\mathbf{s}(t)) - \mu_i \right|, \quad (6)$$

where  $\mathcal{P} = \{\mathcal{P}_1, \dots, \mathcal{P}_B\}$  is a partition of  $\mathcal{S}$  and  $\mu_i$  is the centroid of each cluster  $i$ . The communication policy of ABSA-1 -  $\pi^{c, ABSA-1}(\cdot)$  - is an arbitrary non-injective mapping such that  $\forall k \in \{1, \dots, B\} : \pi^{c, ABSA-1}(\mathbf{s}) = \mathbf{c}^{(k)}$  if and only if  $\mathbf{s} \in \mathcal{P}_k$ . Now let  $C_g$  be a function composition operator such that  $C_g f = g \circ f$ . We define the operator  $\Pi^{m^*} \triangleq C_g$ , with

<sup>7</sup>ABSA's bottleneck arises from the increasing complexity of Q-learning as agents increase in number  $N$ . Similar limitations are in place for any other algorithm that requires a centralized training phase [7], [31]

$$g = \pi^*(\pi^{c, ABSA-1^{-1}}(\cdot))$$

The optimality of the proposed ABSA-1 algorithm is subsequently provided in Theorem 2.

**Lemma 2.** The communication policy  $\pi^{c, ABSA-1}$  - as described by Definition 1 - will carry out lossless compression of observation data w.r.t. the average return if  $|\mathcal{C}| \geq |\mathcal{M}|^N$ .

*Proof.* Appendix B. ■

**Remark:** ABSA-1 will also carry out lossless compression of observation data with respect to the distortion measure introduced in the problem (4). Given the proofs of lemma 2 and lemma 1, the proof of this remark is straightforward and is, therefore, omitted.

The losslessness of quantization in ABSA-1 implies that the  $\pi^{ABSA-1}$  will result in no loss of the system's average return, compared with the case where the optimal policy  $\pi^*(\cdot)$  is used to control the MAS under perfect communications. Consequently, the control policy  $\pi^{m, ABSA-1}(\cdot)$  is optimal. Let us recall once again that here, we do not use a conventional quantization distortion metric, we select a representation of local observation in such a way that the conveyed message maximizes the average task return.

Note that in [7], the authors do not find the higher order function  $\Pi^{m^*}$  that reduces the joint communications and control problem to a task-oriented communication design - instead they solve an approximated version of the task-oriented communication design problem. In this paper, however, we introduce a closed form  $\Pi^{m^*}$  by ABSA-1 that can map every communication policy  $\pi^{c, ABSA-1}$  introduced by ABSA-1, to the exact optimal control policy. This implies that the solutions provided by ABSA-1 are also the optimal solutions to the joint communication and control design (JCCD) problem.

**B. ABSA-2 Algorithm**

We saw earlier in lemma 2 that the communication policy obtained by solving the problem 6 is optimal and can result in a lossless average return performance when  $|\mathcal{C}| \geq |\mathcal{M}|^N$ . To solve the problem 6, however, we need to know  $\pi^*(\mathbf{s}(t))$ . This is a limiting assumption that in ABSA-1 can be translated to two different system models which are less general than the system pictured in Fig. 3: (i) presence of an extra relay between the agents and the central controller where the relay has perfect downlink channels to agents and a single bit-budgeted channel to the CC. (ii) The MAS is only composed of one single agent and a CC where the uplink of the agent is bit-budgeted but its downlink is a perfect channel.

Our second proposed algorithm ABSA-2 removes the need to know  $\pi^*(\mathbf{s}(t))$  and can run under the more general settings shown in Fig. 3. This is done by approximating the local element  $\mathbf{m}_i^*(t)$  of  $\pi^*(\mathbf{s}(t)) = \langle \mathbf{m}_1^*(t), \dots, \mathbf{m}_N^*(t) \rangle$  at

<sup>8</sup>Note that as  $\pi^{c, ABSA-1}(\cdot)$  is non-injective, its inverse would not produce a unique output given any input. Thus, by  $\pi^*(\pi^{c, ABSA-1^{-1}}(\mathbf{c}'))$  we mean  $\pi^*(\mathbf{s}')$ , where  $\mathbf{s}'$  can be any arbitrary output of  $\pi^{c, ABSA-1^{-1}}(\mathbf{c}')$ .

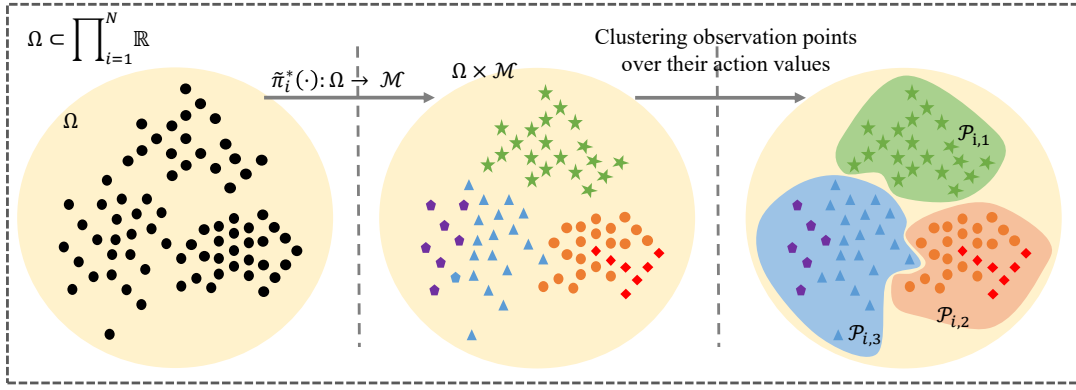


Figure 4. Abstract representation of states in ABSA-2 with  $|\mathcal{C}| = 3$  and  $|\mathcal{M}| = 5$  -  $|\mathcal{M}|$  is represented by the number of shapes selected to show the observation points and  $|\mathcal{C}|$  is represented by the number of clusters shown in the right subplot. The left subplot shows the observation points prior to aggregation. During a centralized training phase we first compute  $\pi^*(\cdot)$  according to which  $\pi_i^*(\cdot) : \Omega \rightarrow \mathcal{M}$  can be obtained. We use the surjection  $\pi_i^*(\cdot)$  to map a high dimensional/precision observation space to a low dimensional/precision space. The middle subplot shows the observation points together with the action values assigned to them - each unique shape represents a unique action value. **This new representation of the observation points, embeds the features of the control problem into the data quantization problem.** Finally, we carry out the clustering of observation points according to their action values - all observation points assigned to (a set of) action values are clustered together. The right subplot shows the aggregated observation space, where all the observation points in each cluster will be represented using the same communication message. The centralized controller which is run using DQN, observes the environment at each time step, through all these aggregated observations/communications it receives from all the agents.

agent  $i$  given the local observation of this agent  $\mathbf{o}_i(t)$ . That is, given a centralized training phase, we will have access to the empirical joint distribution of  $p(\mathbf{o}_i, \mathbf{m}_i^*)$ , using which we can obtain a numerical MAP estimator of  $\hat{\mathbf{m}}_i^*$ . Thus ABSA-2 allows for fully distributed communication policies. In particular, the encoding of the communication messages of each agent is carried out separately by them before they communicate with CC or any other agent. This form of encoding is often referred to as distributed encoding. Furthermore, the encoding carried out by ABSA-2 at each agent is a low-complexity and low-power process that requires no inter-agent communications before hands. In this case, each agent directly communicates its encoded observations to the CC via a bit-budgeted communication channel. In order to improve the learning efficiency at CC, it can take into account all the communications received in the time frame  $[t-d, t]$  to make a control decision  $m(t)$ . Therefore, the ABSA-2 algorithm can strike a trade-off between the complexity of the computations carried out at the CC - directly impacted by the value of  $d$  - and effectiveness of agents' communications - inversely impacted by the value of  $|\mathcal{C}|$ . Moreover, ABSA-2 is straightforwardly extendable to the different values of  $|\mathcal{C}|$  per each agent  $i$ , instead of having only one fixed bit-budget  $R = \log_2 |\mathcal{C}|$  for all agents.

As illustrated in Fig. 4, ABSA-2, each agent  $i$  obtains a communication policy function  $\pi_i^c(\cdot)$  by solving a clustering problem over its local observation space instead of the global state space, formulated as follows:

$$\min_{\mathcal{P}_i} \sum_{j=1}^{|\mathcal{C}|} \sum_{\mathbf{o}_i(t) \in \mathcal{P}_{i,j}} \left| \tilde{\pi}_i^*(\mathbf{o}_i(t)) - \mu_{i,j} \right|, \quad (7)$$

where  $\mathcal{P}_i = \{\mathcal{P}_{i,1}, \dots, \mathcal{P}_{i,|\mathcal{C}|}\}$  is a partition of  $\Omega$ , and

$$\tilde{\pi}_i^*(\mathbf{o}_i(t)) = \operatorname{argmax}_{\mathbf{m}_i^*} p_{\pi^*}(\mathbf{m}_i^* | \mathbf{o}_i(t)), \quad (8)$$

and  $\mathbf{m}_i^*$  is the optimal action of agent  $i$ , which is  $i$ -th element of  $\mathbf{m}^* \triangleq \pi^*(\mathbf{o}_1(t), \dots, \mathbf{o}_N(t))$ . Thus  $\tilde{\pi}_i^*(\mathbf{o}_i(t))$  is the

Algorithm 1. Action Based State Aggregation (ABSA-2)

- 1: **Initialize** replay memory  $D$  to capacity  $10^4 \cdot 000$ .
- 2: **Initialize** state-action value function  $Q(\cdot)$  with random weights  $\theta$ .
- 3: **Initialize** target state-action value function  $Q^t(\cdot)$  with weights  $\theta^t = \theta$ .
- 4: Obtain  $\pi^*(\cdot)$  and  $Q^*(\cdot)$  by solving (2) using Q-learning [41]\*, where  $R \gg H(\mathbf{o}_i(t)) \forall i \in \mathcal{N}$ .
- 5: Compute  $\pi_i^*(\mathbf{o}_i(t)) = \operatorname{Mode}[\mathbf{m}_i^* | \mathbf{o}_i(t)]$ , for  $\forall \mathbf{o}_i(t) \in \Omega$ , for  $i \in \mathcal{N}$ .
- 6: Solve problem (5) by applying k-median clustering to obtain  $\mathcal{P}_i$  and  $\pi_i^c(\cdot)$ , for  $i \in \mathcal{N}$ .
- 7: **for** each episode  $k = 1 : 200 \cdot 000$  **do**
- 8:     Randomly initialize observation  $\mathbf{o}_i(t = 0)$ , for  $i \in \mathcal{N}$
- 9:     Randomly initialize the message  $c(t = 0)$
- 10:    **for**  $t = 1 : T'$  **do**
- 11:       Select  $c_i(t)$ , at agent  $i$ , following  $\pi_i^c(\cdot)$ , for  $i \in \mathcal{N}$
- 12:       Obtain the message  $\langle c_1(t), \dots, c_N(t) \rangle$  at the CC
- 13:       Follow  $\epsilon$ -greedy, at CC, to generate the action  $\mathbf{m}_i(t)$ , for  $i \in \mathcal{N}$
- 14:       Obtain reward  $r(t) = R(s(t), \mathbf{m}(t))$  at the CC
- 15:       Store the transition  $\{c(t), \mathbf{m}(t), r(t), c(t+1)\}$  in  $D$
- 16:        $t \leftarrow t + 1$
- 17:    **end**
- 18:    Sample  $D' = \{c(t'), \mathbf{m}(t'), r(t'), c(t'+1)\}_{t'=t'_1}^{t'=t'_2}$  from  $D$
- 19:    **for** each transition  $t' = t'_1 : t'_2$  of the mini-batch  $D'$  **do**
- 20:       Compute DQN's average loss  $L_{t'}(\theta)$  using (9) },
- 21:       Perform a gradient descent step on  $L_{t'}(\theta)$  w.r.t  $\theta$
- 22:    **end**
- 23:    Update the target network  $Q^t(\cdot)$  every 1000 steps.
- 24: **end**

\*The condition  $R \gg H(\mathbf{o}_i(t)) \forall i \in \mathcal{N}$  mentioned in line 4, implies perfect communications between the agents and the central controller when solving the centralized control problem - for further explanations please refer to, e.g., section III-A.

maximum a posteriori estimator of  $\mathbf{m}_i^* = \pi^*(\mathbf{s}(t))$  given the local observation  $\mathbf{o}_i(t)$ .

Once the clustering in (7) is done, each agent  $i$  will train its local communication policy  $\pi_i^{c, ABSA-2}(\cdot)$ , which is any non-injective mapping such that  $\forall k \in \{1, \dots, |\mathcal{C}|\} : \pi_i^{c, ABSA-2}(\mathbf{o}_i) = \mathbf{c}^{(k)}$  iff  $\mathbf{o}_i \in \mathcal{P}_{i,k}$ . After obtaining the communication policies  $\langle \pi_i^{c, ABSA-2}(\cdot) \rangle_{i=1}^N$ , to obtain a proper control  $\pi^m(\cdot)$  policy at the CC corresponding to the communication policies, we perform a single-agent reinforcement learning. To this end and to manage the complexity of the algorithm for larger values of  $d$ , we propose to use DQN architecture [42] at the CC. When training the DQN at the CC, in we implement the standard DQN, with its standard loss function

$$L_{t'}(\theta) = \frac{1}{2} \left( r(t') + \max_{\mathbf{m}^*} Q^t(c(t'+1), \mathbf{m}^*, \theta^t) - \max_{\mathbf{m}^*} Q(c(t'), \mathbf{m}^*, \theta) \right)^2. \quad (9)$$

#### IV. PERFORMANCE EVALUATION

In this section, we evaluate our proposed schemes via numerical results for the popular multi-agent geometric consensus problem<sup>9</sup>. Through indirect design, ABSA-1 and ABSA-2 never rely on explicit domain knowledge about any specific task, such as geometric consensus. Thus, we conjecture that their indirect design allows them to be applied beyond geometric consensus problems and to a much wider range of tasks. To make the geometric consensus task suitable for the evaluation of our proposed algorithms, similar to [18], we have introduced a bit constraint to the communication channel between the agents and the CC. After evaluating the proposed algorithms in the context of the rendezvous problem, we attempt to explain the behaviour of all the algorithms via the existing metric - positive listening - for measuring the task-effectiveness of communications. As positive listening falls short in explaining all the aspects of the behaviour of the investigated algorithms, we will also introduce a new metric. Called *task relative information*, the new metric assists in further explaining the behaviour of different algorithms with higher accuracy and reliability.

##### A. The geometric consensus problem

Our proposed schemes are evaluated in this section through numerical results for the rendezvous problem [43], [44], which is a specific type of geometric consensus problem under finite observability [29]. Following the instantaneous and synchronous communication model and the star network topology explained in section II-A and Fig. 2 respectively, the rendezvous problem is explained as the following. At each time step  $t$  several events happen in the following order. First, an agent  $i$  obtains a local observation  $\mathbf{o}_i(t)$  - which is equivalent to its own location in the grid world. By following equation (1), the state of the environment  $\mathbf{s}(t)$  at time step  $t$  is defined as the location of all agents in the grid world

<sup>9</sup>In our numerical experiments, the discount factor is assumed to be  $\gamma = 0.9$ . All experiments are done over a grid world of size  $8 \times 8$ , where the goal point of the rendezvous is located at the grid number  $\Omega^T = \{22\}$ .

<sup>10</sup>. Each agent  $i$ , using its quantization/communication policy, generates a compressed version  $c_i(t)$  of its observation to be communicated to the CC via a bit-budgeted communication link. After receiving the quantized observations of all agents, CC follows its control policy to decide and select the joint action vector  $\mathbf{m}(t)$  and communicate each agent  $i$ 's local action  $\mathbf{m}_i(t)$  to it accordingly. The local action  $\mathbf{m}_i(t) \in \mathcal{M}$  that is communicated back to the agent  $i$  via a perfect communication channel is a one-directional move in the greed world, i.e.,  $\mathbf{m}_i(t) \in \mathcal{M} = \{\text{left, right, up, down, pause}\}$ . Given each agent  $i$ 's action  $\mathbf{m}_i(t)$  the environment evolves and transitions to the next time step  $t+1$  where each agent  $i$  obtains a new local observation  $\mathbf{o}_i(t+1)$ . All agents receive a single team reward

$$r_t = \begin{cases} C_1, & \text{if } \exists i, j \in N : \mathbf{o}_i(t) \in \Omega^T \ \& \ \mathbf{o}_j(t) \notin \Omega^T \\ C_2, & \text{if } \nexists i \in N : \mathbf{o}_i(t) \in \Omega - \Omega^T, \\ 0, & \text{otherwise,} \end{cases} \quad (10)$$

where  $C_1 < C_2$  and  $\Omega^T$  is the set of terminal observations i.e., the episode terminates if  $\exists i \in N : \mathbf{o}_i(t) \in \Omega^T$ . Accordingly, when not all agents arrive at the target point, a smaller reward  $C_1 = 1$  is obtained, while the larger reward  $C_2 = 10$  is attained when all agents visit the goal point at the same time. We compare our proposed ABSA algorithms with the heuristic non-communicative (HNC), heuristic optimal communication (HOC) and SAIC algorithms proposed in [18] which are direct schemes to jointly design the communication and control policies for the specific geometric consensus problem solved here. In contrast to ABSA-1 and ABSA-2 which enjoy an indirect design, the direct design of HOC and HNC does not allow them to be applied in any other problem rather than the specific geometric consensus problem with the finite observability i.e., the rendezvous problem explained here.

##### B. Numerical experiment

A constant learning rate  $\alpha = 0.07$  is applied when exact Q-learning is used to obtain  $\pi^*(\cdot)$  and  $\alpha = 0.0007$  when DQN is used to learn  $\pi^m(\cdot)$  for ABSA-2. For the exact Q-learning, a UCB<sup>11</sup> exploration rate of  $c = 1.25$  considered. The deep neural network that approximates the Q-values is considered to be a fully connected feed-forward network with 10 layers of depth, which is optimized using the Adam optimizer. An experience replay buffer of size 10'000 is used with the mini-batch size of 62. The target Q-network is updated every 1000 steps and for the exploration, decaying  $\epsilon$ -greedy with the initial  $\epsilon = 0.05$  and final  $\epsilon = 0.005$  is used [42]. In any figure that the performance of each scheme is reported in terms of the averaged discounted cumulative rewards, the attained rewards throughout training iterations are smoothed using a moving average filter of memory equal to 20,000 iterations. As explained in section III-A, ABSA-1 and ABSA-2 both require

<sup>10</sup>Note that the complete information of the state of the environment  $\mathbf{s}(t)$  at no time step  $t$  is available neither to the CC and nor to an agent  $i$ . Accordingly, agents or the CC have partial access to the information on the state of the environment, to learn more please refer to the section II

<sup>11</sup>UCB is a standard scheme used in exact reinforcement learning to strike a trade-off between the exploration and exploitation [41].



a centralized training phase prior to being capable of being executed in a distributed fashion.

For all black curves, one prior centralized training phase to obtain  $\pi^*(\cdot)$  is required. As detailed in Section III, the proposed algorithms, ABSA-1 and ABSA-2, leverage  $\pi^*(\cdot)$  to design  $\pi^c$  and then  $\pi^m$  afterwards. Dashed curves, HOC and HNC, as proposed by [18] provide heuristic schemes which exploit the domain knowledge of its designer about the rendezvous task making it not applicable to any other task rather than the rendezvous problem. While HOC enjoys a joint control and communication design, HNC runs with no communication. Note that HNC & HOC require communication/coordination between agents prior to the starting point of the task - which is not required for any other scheme. These schemes, introduced by [18], are detailed as the following.

- A joint communication and control policy is designed **using domain knowledge** in the rendezvous problem. HNC agents approach the goal point and wait nearby for a sufficient number of time steps to ensure that the other agent has also arrived. Only after that, they will get to the goal point. Note that this scheme requires communication/coordination between agents prior to the starting point of the task since they have to have agreed upon this scheme of coordination.
- A joint communication and control policy is designed **using domain knowledge** in the rendezvous problem. HOC agents wait next to the goal point until the other agent informs them that they have also arrived there. Only after that, they will get to the goal point. Note that this scheme requires communication/coordination between agents prior to the starting point of the task since they have to have agreed upon this scheme of coordination and communications as well as on the meaning that each communication message entails.

To obtain the results demonstrated in Fig. 5, we have simulated the rendezvous problem for a three-agent system. The black curves illustrate the training phase that is occurring at CC to obtain  $\pi^m$  after  $\pi^c$  is already computed using equations (6) and (7). We observe the lossless performance of ABSA-1 in achieving the optimal average return without requiring any (2nd round) training. To enable fully decentralized quantization of the observation process, ABSA-2 was proposed which is seen to approach the optimal solution as  $d$  grows. All ABSA-2 curves are plotted with  $|\mathcal{C}| = 3$ , and ABSA-1 curve is plotted with  $|\mathcal{C}| = |\mathcal{M}|^N = 125$  in 3 agent scenarios - e.g., in Fig. 5. This figure, however, provides no insights into how each scheme might perform at different rates of quantization, as all the results are reported for a single rate of quantization per scheme.

To see the behaviour of ABSA-2 and other benchmarks HNC, HoC and SAIC at different rates of quantization, Fig. 6 is demonstrated. As expected, with the increase in the size of the quantization codebook, the average return performance of ABSA-2 is gradually improved, such that it approaches near-optimal performance at  $d = 3$ . We also observe the superior performance of ABSA-2 compared with SAIC at very tight

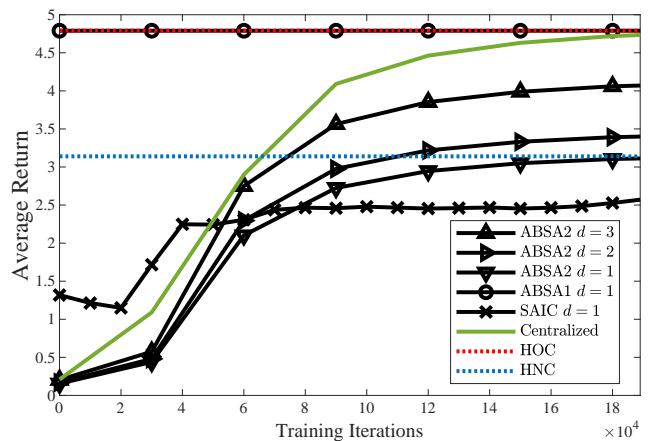


Figure 5. Average return comparison made between the proposed schemes and some benchmarks introduced in [18] - the three agent scenario under constant bit-budget values.

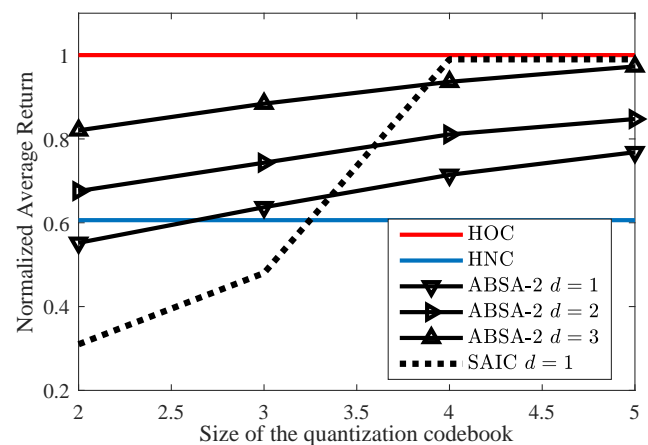


Figure 6. The obtained normalized average return as a function of codebook size  $|\mathcal{C}|$  is compared across a range of schemes: proposed schemes and some benchmarks introduced in [18] - two-agent scenario.

bit-budgets where SAIC's performance sees a drastic drop. It is observed that as  $d$  grows, ABSA-2 approaches the optimal return performance even under higher rates of quantization, however, higher values of  $d$  come at the cost of the increased computational complexity of ABSA-2.

### C. Explainability of the learned communication policies

One common metric to evaluate the effectiveness of communications in the literature [38] is *positive listening*  $I(\mathbf{c}_i(t); \mathbf{m}_j(t))$   $j \in \mathcal{N} - \{i\}$ , which is the mutual information between the communication  $\mathbf{c}_i(t)$  produced by an agent  $i$  and the action  $\mathbf{m}_j(t)$  selected by another agent following the receipt of the communication  $\mathbf{c}_i(t)$  from agent  $i$ . Positive signaling  $I(\mathbf{o}_i(t); \mathbf{c}_i(t))$  is another metric proposed by [38], measuring the mutual information between agent  $i$ 's observation  $\mathbf{o}_i(t)$  and its own produced communication message  $\mathbf{c}_i(t)$  at the same time step. As to be shown below, however, these metrics are unable to fully capture the underlying performance trends of all schemes. Therefore, we, for the first time, introduce a new metric called *task relevant information* (RI) - allowing us to explain the task-effectiveness of the learned communication policies.

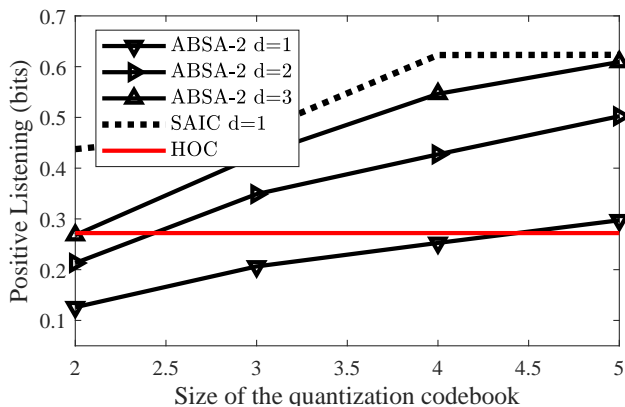


Figure 7. Comparing the positive listening  $I(\mathbf{c}_i(t); \mathbf{m}_j(t))$  performance across a range of schemes.

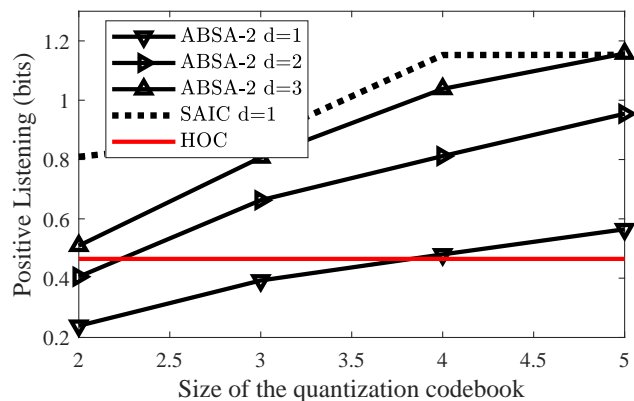


Figure 8. Comparing the positive listening  $I(\mathbf{c}_i(t); \mathbf{m}(t))$  performance across a range of schemes.

Measuring positive listening is one way to quantify the contribution of the communicated messages of agent  $i$  to the action selection of agent  $j$ . Positive signalling, on the other hand, measures the consistency as well as the relevance of the communicated messages  $\mathbf{c}_i(t)$  and the agent's observations  $\mathbf{o}_i(t)$ . As SAIC and ABSA use a deterministic mapping of observation  $\mathbf{o}_i$  to produce the communication message  $\mathbf{c}_i$ , they are always guaranteed to have positive signalling [38] - the degree of which, however, is limited by the uplink channel's bit budget  $R = \log_2 |\mathcal{C}|$ . Thus, among the existing metrics for the measurement of the effectiveness of communications, we limit our numerical studies to the measurement of positive listening. It is known that the higher positive listening is, the stronger (not necessarily better) we expect the coordination between the agents to be. That is, the higher positive listening means higher degree of dependence between agents (their actions and observations) which is not necessarily sufficient for the team agents to fulfill the task.

Figure 7 explains how stronger coordination between agents and the CC is often resulting in an increased performance of the MAS in obtaining a higher average return. For instance, the enhancement in the positive-listening performance of SAIC from  $|\mathcal{C}| = 3$  to  $|\mathcal{C}| = 4$  quantizer in Fig. 7 is resulting in an improved average return performance, as shown in Fig. 6. This metric also reasonably explains the enhancement of ABSA-2 performance in obtaining higher return by increasing  $d$  - the memory of the CC - and the size of the quantization codebook  $|\mathcal{C}|$ . Moreover, stronger coordination between agents and CC is visible in ABSA-2 when compared with HOC. Thus, we expect better average return performance for ABSA-2 which is in contrast to the results of Fig. 5. This event suggests that stronger coordination - measured by positive listening - may not necessarily result in an improved average return performance as the coordination may not be perfectly aligned with task needs.

The curve concerning the HOC scheme allows us to recall that a positive listening of 0.3 (bit) is sufficient to maintain the coordination required for optimal performance in the aforementioned geometric consensus task. Therefore, in the ABSA-2 and SAIC schemes, there is still an unnecessary influence from the side of the communication messages to the actions

selected by the receiving end. In fact, not all the information received from the receiving end has contributed to the higher average return of the system. Accordingly, there is yet, some unnecessary data in the communication messages designed by ABSA that contain no task-specific/useful information.

Thus we believe that positive listening cannot explicitly quantify the effectiveness of the task-oriented communication algorithms; therefore they fall short in explaining the behaviour of these algorithms. Even when positive listening is computed as  $I(\mathbf{c}_i(t); \mathbf{m}(t))$  to capture the mutual information between the communication of agent  $i$  and the control signals of all agents we arrive at almost similar patterns - Fig. 8.

Figure 9, investigates the performance of multiple schemes via a novel performance metric: task-relevant information (TRI). Here we define the task-relevant information metric to be

$$I(\pi^c(\mathbf{o}_i(t)); \pi^*(\mathbf{s}(t))) = I(\mathbf{c}_i(t); \mathbf{m}^*(t)), \quad (11)$$

which measures the mutual information (in bits) between the communicated message of agent  $i$  and the vector  $\mathbf{m}^*(t)$  of joint optimal actions at the CC - which is selected by the optimal centralized control policy  $\pi^*(\cdot)$ . As demonstrated by Fig. 9, TRI is an indirect metric of the effectiveness of communications that can explain the behaviour of different communication designs. It is also observed that the TRI metric magnifies the performance gap between different schemes as they get closer to the optimal performance. Nevertheless, TRI can be utilized as a standalone measure to quantify the effectiveness of a communication design since it almost perfectly predicts the average return performance of the a communication policy - without the need for the communication to be tested when solving the real task.

Note that, we measure the task-effectiveness of a quantization algorithm based on the average return that can be obtained when using it. Further, to measure the average return that can be obtained under the communication policies  $\langle \pi_1^c(\cdot), \dots, \pi_N^c(\cdot) \rangle$ , we have to design the control policy  $\pi^m(\cdot)$  at the CC that selects the control vector  $\mathbf{m}(t)$  having access to only the quantized observations of the agents  $\mathbf{c}(t)$ . Accordingly, we cannot measure the effectiveness of the

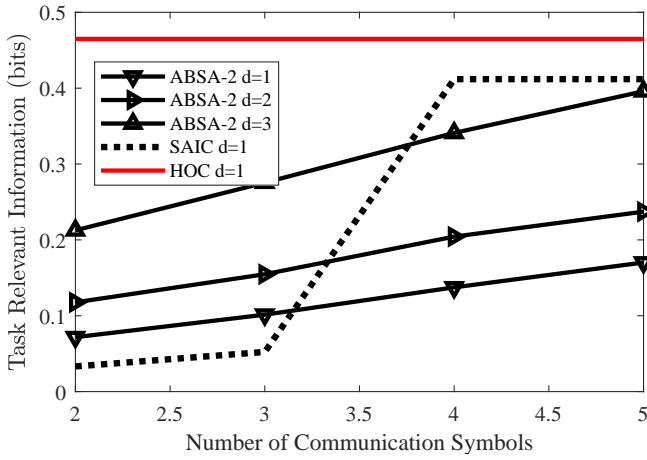


Figure 9. Comparing the task-relevant information (TRI) performance across a range of schemes. It is observed that TRI can comprehensively explain the behaviour of all task-effective quantization schemes in a certain task without the need to measure their effectiveness via their resulting average return in the task - compare this figure with Fig. 6 .

communication policy of an MAS without having a specific design for their control policy. Even after the design of the control policy of the MAS, it is challenging to understand if the suboptimal performance of the algorithm is caused by an ineffective design of the control policy or the communication policy. In fact, it is hard disentangle the effect of the control and communication policies on the MAS's average return. Our proposed metric TRI can facilitate measuring the performance of any communication policy in isolation and without the effect of the control policy being present in the numerical values of TRI.

Accordingly, the importance of introducing this metric is multi-fold: (i) by using TRI as an indirect metric we can measure the effectiveness of a communication policy for any specific task; (ii) it allows us to measure the effectiveness of the communication scheme prior to the design of any control policy; (iii) it helps to design task effective communication policies in complete separation from the control policy design.

## V. CONCLUSION

In this paper, we have investigated the joint design of control and communications in a MAS under centralized control and distributed communication policies. We first proposed an action-based state aggregation algorithm (ABSA-1) for lossless compression and provided analytical proof of its optimality. Then we proposed ABSA-2, which offers a fully distributed communication policy and can trade computational complexity for communication efficiency. We finally demonstrated the task-effectiveness of the proposed algorithms via numerical experiments performed on a geometric consensus problem via a number of representative metrics. Furthermore, our numerical studies demonstrate the pressing need for further research on finding a metric that can measure/explain the task-effectiveness of communications with more accuracy. And, scalability in task-oriented design is yet another central challenge to be addressed in future research.

## APPENDIX A PROOF OF LEMMA 1

*Proof.* Applying Adam's law on equation (3) yields

$$\operatorname{argmax}_{\pi} \mathbb{E}_{p(c(t))} \left\{ \mathbb{E}_{p_{\pi^c, \pi^m}(\{tr\}_{t'}^{T'} | c(t))} \{g(t') | c(t)\} \right\}, \text{ s.t. } |\mathcal{C}| \leq 2^R \quad (12)$$

where  $c(t)$  is generated by the communication policy  $\pi^c$  and the joint pmf of the system's trajectory  $\{tr\}_{t'}^{T'}$  is directly influenced by the action policy  $\pi^m$ . The conditional pmf  $p_{\pi^c, \pi^m}(\{tr\}_{t'}^{T'} | c(t))$  is the joint probability of the trajectory of the system given the received communication  $c(t)$  when policies  $\pi^c(\cdot)$  and  $\pi^m(\cdot)$  are followed. We proceed by negating the equation (12) and adding a second term to the objective function which is constant with respect to the decision variables of the problem to have

$$\operatorname{argmin}_{\pi^c} \mathbb{E}_{p(s(t))} \left\{ \mathbb{E}_{p_{\pi^*}(\{tr\}_{t'}^{T'} | s(t))} \{g(t') | s(t)\} \right\} - \mathbb{E}_{p(c(t))} \left\{ \mathbb{E}_{p_{\pi^c, \pi^m}(\{tr\}_{t'}^{T'} | c(t))} \{g(t') | c(t)\} \right\}, \text{ s.t. } |\mathcal{C}| \leq 2^R. \quad (13)$$

We replace the conditional expectation of system return by the value function  $V(\cdot)$ , [41](Ch. 3.5), and we will have

$$\operatorname{argmin}_{\pi^c} \mathbb{E}_{p(s(t))} \left\{ V^{\pi^*}(s(t)) \right\} - \mathbb{E}_{p(c(t))} \left\{ V^{\pi^m}(c(t)) \right\}, \text{ s.t. } |\mathcal{C}| \leq 2^R. \quad (14)$$

Note that the empirical joint distribution of  $c(t)$  can be obtained by following the communication policy  $\pi^c$  on the empirical distribution of  $s(t)$ .

$$\operatorname{argmin}_{\pi^c} \mathbb{E}_{p(s(t))} \left\{ V^{\pi^*}(s(t)) \right\} - \mathbb{E}_{p(s(t))} \left\{ V^{\pi^m}(c(t)) \right\}, \text{ s.t. } |\mathcal{C}| \leq 2^R. \quad (15)$$

As  $V^{\pi^*}(s(t)) - V^{\pi^m}(c(t)) \geq 0$  is true for any  $s(t) \in \mathcal{S}$ , merging the two expectations results in

$$\operatorname{argmin}_{\pi^c} \mathbb{E}_{p(s(t))} \left| V^{\pi^*}(s(t)) - V^{\pi^m}(c(t)) \right|, \text{ s.t. } |\mathcal{C}| \leq 2^R, \quad (16)$$

which concludes the proof of the lemma.  $\blacksquare$

## APPENDIX B PROOF OF LEMMA 2

*Proof.* We depart from the result of lemma 1 - problem (4). By taking the expectation over the empirical distribution of  $s(t)$  and applying Bellman optimality equation, we obtain

$$\operatorname{argmin}_{\pi} \frac{1}{n} \sum_{t=1}^n \left| Q^{\pi^*}(s(t), \pi^*(s(t))) - Q^{\pi^m}(c(t), \pi^m(\pi^c(s(t)))) \right|, \text{ s.t. } |\mathcal{C}| \leq 2^R, \quad (17)$$

where the vector  $\pi^c(s(t))$  is of  $N$  dimensions and its  $i$ -th element is  $c_i(t)$ . We proceed by plugging  $\pi^{c, ABSA-1}(\cdot)$  and  $\Pi^{m^*}$ , according to the definition 1, into the equation (17) to obtain

$$\frac{1}{n} \sum_{t=1}^n \left| Q^{\pi^*}(s(t), \pi^*(s(t))) - Q^{\pi^*}(c(t), \pi^*(s')) \right|, \quad (18)$$

where  $s' = \pi^{c, ABSA-1}(\pi^{c, ABSA-1}(s(t)))$ , and any possible value for it lies in the same subset  $\mathcal{P}_{k'}$  as  $s(t)$  does, while according to the definition of  $\mathcal{P}_{k'}$ , we know  $\pi^*(s(t)) = \pi^*(s')$ ,

if  $|\mathcal{C}| \geq |\mathcal{M}|^N$ . Thus, by replacing  $\pi^*(s')$  in with  $\pi^*(s(t))$  in equation (19) we get

$$\frac{1}{n} \sum_{t=1}^n \left| Q^{\pi^*}(s(t), \pi^*(s(t))) - Q^{\pi^*}(s(t), \pi^*(s(t))) \right| = 0. \quad (19)$$

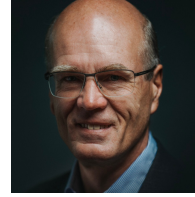
This concludes the proof of theorem 2. ■

## REFERENCES

- [1] L. S. Vailshery, "Number of internet of things (iot) connected devices worldwide from 2019 to 2021, with forecasts from 2022 to 2030," Aug 2022. [Online]. Available: <https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/>
- [2] B. Güler, A. Yener, and A. Swami, "The semantic communication game," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 4, pp. 787–802, 2018.
- [3] H. Tong, Z. Yang, S. Wang, Y. Hu, W. Saad, and C. Yin, "Federated learning based audio semantic communication over wireless networks," in *2021 IEEE Global Communications Conference (GLOBECOM)*, 2021, pp. 1–6.
- [4] N. Pappas and M. Kountouris, "Goal-oriented communication for real-time tracking in autonomous systems," in *2021 IEEE International Conference on Autonomous Systems (ICAS)*, 2021, pp. 1–5.
- [5] E. Calvanese Strinati and S. Barbarossa, "6g networks: Beyond shannon towards semantic and goal-oriented communications," *Computer Networks*, vol. 190, p. 107930, 2021.
- [6] 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.
- [7] 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.
- [8] C. E. Shannon and W. Weaver, "The mathematical theory of communication [1949]. urbana, il," 1959.
- [9] L. Hu, G. Wu, Y. Xing, and F. Wang, "Things2vec: Semantic modeling in the internet of things with graph representation learning," *IEEE Internet of Things Journal*, vol. 7, no. 3, pp. 1939–1948, 2020.
- [10] J. Cai, W. Zhong, and J. Luo, "Seminer: Side-information-based semantics miner for proprietary industrial control protocols," *IEEE Internet of Things Journal*, vol. 9, no. 22, pp. 22 796–22 810, 2022.
- [11] T.-Y. Tung, S. Kobus, J. P. Roig, and D. Gündüz, "Effective communications: A joint learning and communication framework for multi-agent reinforcement learning over noisy channels," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 8, pp. 2590–2603, 2021.
- [12] M. P. Mota, A. Valcarce, J.-M. Gorce, and J. Hoydis, "The emergence of wireless mac protocols with multi-agent reinforcement learning," *arXiv preprint arXiv:2108.07144*, 2021.
- [13] N. Shlezinger and Y. C. Eldar, "Deep task-based quantization," *Entropy*, vol. 23, no. 1, p. 104, 2021.
- [14] M. A. Gutierrez-Estevéz, Y. Wu, and C. Zhou, "Learning to communicate with intent: An introduction," *arXiv preprint arXiv:2211.09613*, 2022.
- [15] C. Zhang, H. Zou, S. Lasaulce, W. Saad, M. Kountouris, and M. Bennis, "Goal-oriented communications for the iot and application to data compression," *IEEE Internet of Things Magazine*, vol. 5, no. 4, pp. 58–63, 2022.
- [16] N. Shlezinger and Y. C. Eldar, "Task-based quantization with application to mimo receivers," *arXiv preprint arXiv:2002.04290*, 2020.
- [17] 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.
- [18] A. Mostaani, T. X. Vu, S. Chatzinotas, and B. Ottersten, "Task-oriented data compression for multi-agent communications over bit-budgeted channels," *IEEE Open Journal of the Communications Society*, vol. 3, pp. 1867–1886, 2022.
- [19] M. Kountouris and N. Pappas, "Semantics-empowered communication for networked intelligent systems," *IEEE Communications Magazine*, vol. 59, no. 6, pp. 96–102, 2021.
- [20] R. Carnap, Y. Bar-Hillel *et al.*, "An outline of a theory of semantic information," 1952.
- [21] H. Zhang, S. Shao, M. Tao, X. Bi, and K. B. Letaief, "Deep learning-enabled semantic communication systems with task-unaware transmitter and dynamic data," *arXiv preprint arXiv:2205.00271*, 2022.
- [22] P. A. Stavrou and M. Kountouris, "A rate distortion approach to goal-oriented communication," in *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022, pp. 590–595.
- [23] 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.
- [24] 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.
- [25] J. Liu, S. Shao, W. Zhang, and H. V. Poor, "An indirect rate-distortion characterization for semantic sources: General model and the case of gaussian observation," *arXiv preprint arXiv:2201.12477*, 2022.
- [26] C.-M. Chou, C.-Y. Li, W.-M. Chien, and K.-c. Lan, "A feasibility study on vehicle-to-infrastructure communication: Wifi vs. wimax," in *2009 tenth international conference on mobile data management: systems, services and middleware*. IEEE, 2009, pp. 397–398.
- [27] Y.-C. Liu, J. Tian, C.-Y. Ma, N. Glaser, C.-W. Kuo, and Z. Kira, "Who2com: Collaborative perception via learnable handshake communication," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6876–6883.
- [28] H. Zhu, Q. Wu, X.-J. Wu, Q. Fan, P. Fan, and J. Wang, "Decentralized power allocation for mimo-noma vehicular edge computing based on deep reinforcement learning," *IEEE Internet of Things Journal*, vol. 9, no. 14, pp. 12 770–12 782, 2022.
- [29] A. Barel, R. Manor, and A. M. Bruckstein, "Come together: Multi-agent geometric consensus," *arXiv preprint arXiv:1902.01455*, 2017.
- [30] S. Tatikonda and S. Mitter, "Control under communication constraints," *IEEE Transactions on automatic control*, vol. 49, no. 7, pp. 1056–1068, 2004.
- [31] 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.
- [32] F. A. Oliehoek, C. Amato *et al.*, *A concise introduction to decentralized POMDPs*. Springer, 2016, vol. 1.
- [33] Z. Ding, W. Hong, L. Zhu, T. Huang, and Z. Lu, "Sequential communication in multi-agent reinforcement learning," 2021.
- [34] J. Albowicz, A. Chen, and L. Zhang, "Recursive position estimation in sensor networks," in *Proceedings Ninth International Conference on Network Protocols. ICNP 2001*. IEEE, 2001, pp. 35–41.
- [35] S. Dorvash and S. Pakzad, "Stochastic iterative modal identification algorithm and application in wireless sensor networks," *Structural Control and Health Monitoring*, vol. 20, no. 8, pp. 1121–1137, 2013.
- [36] 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.
- [37] 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.
- [38] 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.
- [39] L. Li, T. J. Walsh, and M. L. Littman, "Towards a unified theory of state abstraction for mdps," in *AI&M*, 2006.
- [40] A. K. McCallum, *Reinforcement learning with selective perception and hidden state*. University of Rochester, 1996.
- [41] R. S. Sutton and A. G. Barto, *Introduction to reinforcement learning*, 2nd ed. MIT Press, Nov. 2017, vol. 135.
- [42] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [43] 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>
- [44] 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.



**Arsham Mostaani** (Student Member, IEEE) received his Ph.D. degree in information systems from the University of Luxembourg (Luxembourg), in 2023. He has gained working experience as a researcher at the King's College London as well as Nokia Bell Labs - Stuttgart. He is now an associate research fellow at the Centre for the Security, Reliability and Trust (SnT) and also a visiting researcher at Eurecom, France. His research interests include task-oriented communications, semantic communications, multiagent learning, and distributed systems.



**Björn Ottersten** (Fellow Member, IEEE) has received the M.Sc. degree in electrical engineering and applied physics from Linköping University, Linköping, Sweden, in 1986, and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 1990. He has held research positions with the Department of Electrical Engineering, Linköping University, the Information Systems Laboratory, Stanford University, the Katholieke Universiteit Leuven, Leuven, Belgium, and the University of Luxembourg, Luxembourg. From 1996 to

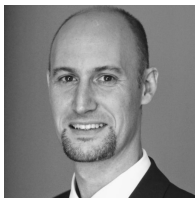
1997, he was the Director of Research with ArrayComm, Inc., a start-up in San Jose, CA, USA, based on his patented technology. In 1991, he was appointed Professor of signal processing with the Royal Institute of Technology (KTH), Stockholm, Sweden. Dr. Ottersten has been Head of the Department for Signals, Sensors, and Systems, KTH, and Dean of the School of Electrical Engineering, KTH. He is the founding Director for the Interdisciplinary Centre for Security, Reliability and Trust, University of Luxembourg. He is a recipient of the IEEE Signal Processing Society Technical Achievement Award, the EURASIP Group Technical Achievement Award, and the European Research Council (ERC) advanced research grant twice. He has co-authored journal papers that received the IEEE Signal Processing Society Best Paper Award in 1993, 2001, 2006, 2013, and 2019, and 9 IEEE conference papers best paper awards. He has been a board member of IEEE Signal Processing Society, the Swedish Research Council and currently serves of the boards of EURASIP and the Swedish Foundation for Strategic Research as well as on the ERC Scientific Council. Dr. Ottersten has served as Editor in Chief of EURASIP Signal Processing, and acted on the editorial boards of IEEE Transactions on Signal Processing, IEEE Signal Processing Magazine, IEEE Open Journal for Signal Processing, EURASIP Journal of Advances in Signal Processing and Foundations and Trends in Signal Processing. He is a fellow of IEEE, EURASIP, and AAlA.



**Thang X. Vu** (Senior Member, IEEE) received the B.S. and the M.Sc., both in Electronics and Telecommunications Engineering, from the VNU University of Engineering and Technology, Vietnam, in 2007 and 2009, respectively, and the Ph.D. in Electrical Engineering from the University Paris-Sud, France, in 2014. In 2010, he received the Allocation de Recherche fellowship to study Ph.D. in France.

From July 2014 to January 2016, he was a post-doctoral researcher with the Singapore University of Technology and Design (SUTD), Singapore. Currently,

he is a research scientist at the Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg. His research interests are in the field of wireless communications, with particular interests of applications of optimization and machine learning on design and analyze the multi-layer 6G networks. He was a recipient of the SigTelCom 2019 best paper award. Currently, he is serving as an Associate Editor of the IEEE Communications Letters. He is a senior member of the IEEE.



**Symeon Chatzinotas** (Fellow Member, IEEE) is currently Full Professor / Chief Scientist I and Head of the SIGCOM Research Group at SnT, University of Luxembourg. In the past, he has been a Visiting Professor at the University of Parma, Italy and he was involved in numerous Research and Development projects for the National Center for Scientific Research Demokritos, the Center of Research and Technology Hellas and the Center of Communication Systems Research, University of Surrey. He received the M.Eng. degree in telecommunications

from the Aristotle University of Thessaloniki, Thessaloniki, Greece, in 2003, and the M.Sc. and Ph.D. degrees in electronic engineering from the University of Surrey, Surrey, U.K., in 2006 and 2009, respectively. He was a co-recipient of the 2014 IEEE Distinguished Contributions to Satellite Communications Award, the CROWNCOM 2015 Best Paper Award and the 2018 EURASIP JWCN Best Paper Award. He has (co-)authored more than 400 technical papers in refereed international journals, conferences and scientific books. He is currently in the editorial board of the IEEE Open Journal of Vehicular Technology and the International Journal of Satellite Communications and Networking.