Centralized Control of a Multi-Agent System Via Distributed and Bit-Budgeted Communications

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

***Abstract*—We consider a distributed quantization problem that arises when multiple edge devices, i.e., 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, given its local processing resources, 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 (MAS). 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 an MAS. We evaluate the algorithms - with average return as the performance metric - using numerical experiments performed to solve a multi-agent geometric consensus problem.**

***Index Terms*—Task-oriented data compression, distributed edge processing, communications for machine learning, multi-agent systems, semantic communications.**

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] and the timeliness of information [3]. 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 [4]. 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 communications towards the eventual task-effectiveness [4].

According to Shannon and Weaver, communication problems can be divided into three levels [5]: (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] as well as the effectiveness problem [4], [6]–[10] have attracted extensive research interest.

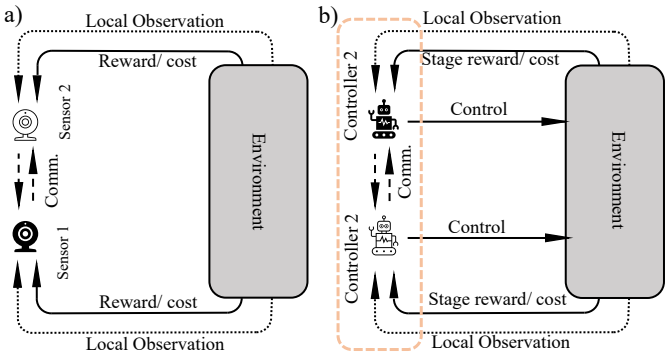

Figure 1. Task-effective communications for a) an estimation vs. b) a control task at the edge - the orange dashed box is detailed in and Fig. 2.

The focus of this work is on the effectiveness problem where in contrast to semantic level and technical level communication design, the performance of the communication system is ultimately measured in terms of the average return/cost linked to the task [7]. 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 [3], [9] or a control task [6]–[8], [10].

There are fundamental differences between the design of task-effective communications for an estimation vs. a control task at the edge - Fig. 1. (i) In the latter, each edge device i.e., an 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 - [6], [7], [10]. In the estimation tasks, however, the source of information is often assumed to be an i.i.d. stochastic process [9]. (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. (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. The considered system architecture can find applications in several domains such as Internet of Things, emerging cyber-physical systems, real-time interactive systems, vehicle-to-infrastructure communication and collaborative perception.

We consider a novel problem setting in which an MAS is controlled via a central controller who has access to agents' local observations only through bit-budgeted distributed communications. This problem setting finds applications in collaboration perception systems as well as vehicle-to-infrastructure communications, which cannot been addressed by the problem settings investigated in the prior similar art. 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. Moreover, our analytical studies help us to craft an indirect [1] task-effective data quantization algorithm - ABSA-2. 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 - making a proper choice for edge processing applications with low to high processing power available at the edge. Finally, 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 [6], in a multi-agent scenario with local observability and bit-budgeted communications.

[1]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 [9]. As defined by [4]: "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".

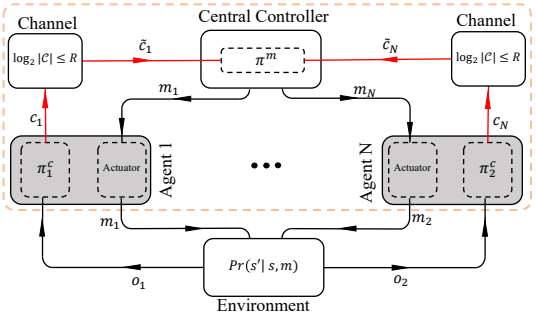

Figure 2. Illustration of the interactions of the CC and agents for the control of the environment. The red link shows the communication channels that are bit-budgeted - implying the incomplete observability of the CC.

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. Bold font is used for matrices or scalars which are random and their realizations follow simple font.

II. Problem Statement

We consider a multi-agent system in which multiple agents $i \in \mathcal{N} = \{1, 2, ..., N\}$ collaboratively solve a task with the aid of a CC, as shown in Fig. 2. 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. The considered setting is similar to conventional centralized control of multi-agent systems [11], except for the fact that the communications from the agents to the CC are transmitted over a bit-budgeted communication channel. We note that there is no direct inter-agent communication in the considered system - communications ocurr only between agents and the CC. 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 $\mathbf{s}(t) \triangleq \langle \mathbf{o}_1(t), \ldots, \mathbf{o}_N(t) \rangle$. 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 $\mathbf{m}(t) \triangleq \langle \mathbf{m}_1(t), ..., \mathbf{m}_N(t) \rangle$. The observation space Ω, state-space $\mathcal{S}$, and action space $\mathcal{M}$ are all discrete sets. The environment is governed by an underlying[2] Markov Decision Process that is described by the tuple $M = \{\mathcal{S}, \mathcal{M}^N, r(\cdot), \gamma, T(\cdot)\}$, where $r(\cdot) : \mathcal{S} \times \mathcal{M}^N \to \mathbb{R}$ is the per-stage reward function and the scalar $0 \le \gamma \le 1$ is the discount factor. The function $T(\cdot) : \mathcal{S} \times \mathcal{M}^N \times \mathcal{S} \to [0, 1]$ is a conditional probability mass function (PMF) which represents state transitions such that $T\big(\mathsf{s}(t+1), \mathsf{s}(t), \mathsf{m}(t)\big) = \Pr\big(\mathsf{s}(t+1) | \mathsf{s}(t), \mathsf{m}(t)\big)$. According

[2]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.

to the per-stage reward signals, the system's return within the time horizon T' is denoted by

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

While the system state is jointly observable by the agents, each agent i's observation $\mathbf{o}_i(t)$ is local[3]. Once per time step, agent $i \in \mathcal{N}$ is allowed to transmit its local observations through a communication message $\mathbf{c}_i(t)$ to the CC. The communications between agents and the CC are done in a synchronous (not sequential) and simultaneous (not delayed) fashion [10]. Each agent i generates its communication message $\mathbf{c}_i(t)$ by following its communication policy $\pi_i^c(\cdot) : \Omega \to \mathcal{C}$. 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 $\mathbf{c}_i(t)$ which will be received by the CC in the same time-step t. The code-book $\mathcal{C}$ is a set composed of a finite number of communication symbols $\mathsf{c}, \mathsf{c}', \mathsf{c}'', ..., \mathsf{c}^{(|\mathcal{C}|-1)}$ - 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 cardinality of the communication symbol space should follow the inequality $|\mathcal{C}| \leq 2^R$. The CC exploits the received communication messages $\mathbf{c}(t) \triangleq \langle \mathbf{c}_1(t), ..., \mathbf{c}_N(t) \rangle$ within the last d number of time-steps to generate the action signal $\mathsf{m}(t)$ following the control policy $\pi^m(\cdot) : \mathcal{C}^{Nd} \to \mathcal{M}^N$. 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.

Now we define the joint control and communication design (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 $\pi^c \triangleq \langle \pi_1^c(\cdot), ..., \pi_N^c(\cdot) \rangle$ to solve the following variational dynamic programming

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

where the expectation is taken over the joint PMF of system's trajectory $\{\mathsf{tr}\}_{t'}^{T'} = \mathsf{o}_1(t'), ..., \mathsf{o}_N(t'), \mathsf{m}(t'), ..., \mathsf{o}_1(T'), ..., \mathsf{o}_N(T'), \mathsf{m}(T')$, when the agents follow the policy tuple π. In the next section, similar to [4] we will disentangle the design of action and communication policies for a CC aided multi-agent system via action-based quantization of observations.

III. Action-based Lossless compression of observations

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 (2) can also be expressed as a generalized data quantization problem as follows*

$$\underset{\pi}{\textit{argmin}} \quad \mathbb{E}_{p(\mathbf{s}(t))} \Big| V^{\pi^*}\big(\mathbf{s}(t)\big) - V^{\pi^m}\big(\mathbf{c}(t)\big) \Big|, \;\; \textit{s.t.} \;\; |\mathcal{C}| \leq 2^R, \quad (3)$$

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

Proof. Due to the space limit, we have removed the proof which will be available in the extended version of the paper. ■

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 (3) is not conventional. In fact, the variational minimization problem is solved over the vector space of joint quantization policies π^c and action policy π^m functions.

As mentioned earlier, in this paper we try to set yet another novel example - in addition to [10] - for the use of a generic framework to solve JCCD problem.

In [6], 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 Π^{m^*}. In particular, the functional Π^{m^*} is a map from the vector space $\mathcal{K}^c$ of all possible communication policies π^c to the vector space $\mathcal{K}^m$ of optimal corresponding control policy $\pi^{m^*}(\cdot)$. Upon the availability of the functional Π^{m^*}, wherever the function π^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 π^c are to be designed. While in [6], authors use an approximation of $\Pi^{m^*}(\pi^c)$ to obtain a task-oriented quantizer design problem, in ABSA-1 we derive an exact solution for a simplified version of (3) - where a relay exists between the agents and the central controller. To adapt ABSA-1 to the generic setting of problem (3), in ABSA-2, we will then need to replace the output of the optimal action policy function with its maximum a posteriori (MAP) estimator.

A. ABSA-1 Algorithm

In the proposed ABSA-1, we assume that the agents communicate with the CC via the aid of a relay. Although the relay has full access to the agents' communication messages, i.e., $\mathbf{c}_i, \forall i$, the relay-CC channel is bit-budgeted. Such assumption is useful to facilitate our analytical studies on the problem (3), allows to establish theoretical proof of 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

[3]In our problem setting, each single agent do 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 reinforcement learning algorithms.

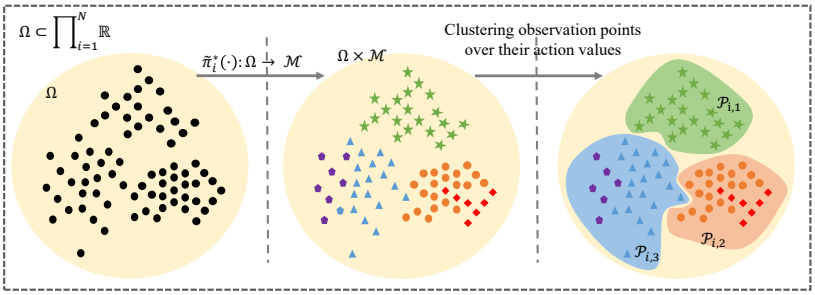

Figure 3. Abstract representation of states in ABSA-2 with $|\mathcal{C}| = 3$ and $|\mathcal{M}| = 5$ - $|\mathcal{M}|$ is represented by the number of shapes selected to show the observation points and $|\mathcal{C}|$ is represented by the number of clusters shown in the right subplot. Left subplot shows the observation points prior to aggregation. During a centralized training phase we first compute $\pi^*(\cdot)$ according to which $\pi_i^*(\cdot) : \Omega \to \mathcal{M}$ can be obtained. We use the surjection $\pi_i^*(\cdot)$ to map a high dimensional/precision observation space to a low dimensional/precision space. 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 source coding 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.

$\mathbf{c}$ iff $\pi^*\big(\mathbf{s}^{(i)}\big) = \pi^*\big(\mathbf{s}^{(j)}\big)$, where $\pi^*(\cdot) : \mathcal{S} \to \mathcal{M}^N$ is the optimal control policy of the agents, given the access of observations from all agents. Thus, similar to the SAIC algorithm introduced in [6], 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 $\pi^*(\cdot)$; (ii) solving the task-oriented data quantization problem to find π^c via a form of data clustering; (iii) finding the π^m corresponding to π^c.

In order to explain ABSA-1, we introduce the problem of task-based information compression with centralized control (TBIC-CC). TBIC-CC 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 TBIC-CC problem from the JCCD problem, we use the functional Π^{m^*} to replace $\pi^m(\cdot)$ with $\Pi^{m^*}\big(\pi^c\big)$. Upon the availability of Π^{m^*}, by plugging it into the JCCD problem (2), we will have a new problem

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

where we maximize the system's return with respect to only the communication policy $\pi^c(\cdot)$ of the local relay. The optimal control policy $\pi^{m^*}(\cdot)$ of the CC is automatically computed by the mapping $\Pi^{m^*}\big(\pi^c(\cdot)\big)$. The problem is called here as the TBIC-CC problem. Upon the availability of Π^{m^*}, the JCCD problem (2) can be reduced to TBIC-CC. Definition 1 is provided to formalize a precise approach to solve TBIC-CC via obtaining the communication policy of the relay $\pi^c(\cdot)$ as well as the corresponding Π^{m^*}, to solve (2).

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

$$\min_{\mathcal{P}} \quad \sum\nolimits_{i=1}^{|\mathcal{C}|} \sum\nolimits_{\mathbf{s}(t)\in\mathcal{P}_i} \Big|\pi^*\big(\mathbf{s}(t)\big) - \mu_i\Big|, \qquad (4)$$

where $\mathcal{P} = \{\mathcal{P}_1, ..., \mathcal{P}_B\}$ is a partition of $\mathcal{S}$ and μ_i is the centroid of each cluster i. We define $\pi^{c,ABSA-1}(\cdot)$ to be any non-injective mapping such that $\forall k \in \{1, ..., B\}$: $\pi^{c,ABSA-1}(\mathbf{s}) = \mathbf{c}^{(k)}$ 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 $g = \pi^*\big(\pi^{c,ABSA-1^{-1}}(\cdot)\big)$[4] .

The optimality of the proposed ABSA-1 algorithm is provided in Lemma 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. Due to the space limit, we have removed the proof which will be available in the extended version of the paper. ■

The losslessness of quantization in ABSA-1 implies that the π^{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 multi-agent system under perfect communications. Consequently, the control policy $\pi^{m,ABSA-1}(\cdot)$ is optimal. Let us recall that, we do not use a conventional quantization distortion metric, we select a representation of local observation in such a way that the conveyed message maximize the average task return.

B. ABSA-2 Algorithm

Our second proposed algorithm ABSA-2 removes the need for the presence of the relay, thus allowing 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

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

channel. In order to improve the learning efficiency at the 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, 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 fiexed bit-budget $|\mathcal{C}|$ for all agents. ABSA-2 is detailed in Algorithm 1.

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

Initialize replay memory D to capacity 10'000.
Initialize state-action value function $Q(\cdot)$ with random θ.
Initialize target state-action value function $Q^t(\cdot)$ with weights $\theta^t=\theta$.
Obtain $\pi^*(\cdot)$ and $Q^*(\cdot)$ by solving (2) using Q-learning [12]*, where $R >> H(\mathbf{o}_i(t))\ \forall i\in\mathcal{N}$.
Compute $\pi_i^*(\mathsf{o}_i(t)) = \text{Mode}\big[\mathbf{m}_i^*|\mathsf{o}_i(t)\big]$, for $\forall \mathsf{o}_i(t)\in\Omega$, for $i\in\mathcal{N}$.
Solve problem (5) by applying k-median clustering to obtain $\mathcal{P}_i$ and $\pi_i^c(\cdot)$, for $i\in\mathcal{N}$.
for each episode $k=1:200$'000 **do**
Randomly initialize observation $\mathsf{o}_i(t=0)$, for $i\in\mathcal{N}$
Randomly initialize the message $\mathsf{c}(t=0)$
for $t=1:T'$ **do**
Select $\mathsf{c}_i(t)$, at agent i, following $\pi_i^c(\cdot)$, for $i\in\mathcal{N}$
Obtain the message $\langle \mathsf{c}_1(t),...,\mathsf{c}_N(t)\rangle$ at the CC
Follow ϵ-greedy, at CC, to generate the action $\mathsf{m}_i(t)$, for $i\in\mathcal{N}$
Obtain reward $r(t)=R\big(\mathsf{s}(t),\mathsf{m}(t)\big)$ at the CC
Store the transition $\Big\{\mathsf{c}(t),\mathsf{m}(t),r(t),\mathsf{c}(t+1)\Big\}$ in D
$t\leftarrow t+1$
end
Sample $D'=\Big\{\mathsf{c}(t'),\mathsf{m}(t'),r(t'),\mathsf{c}(t'+1)\Big\}_{t'=t'_1}^{t'=t'_{62}}$ from D
for each transition $t'=t'_1:t'_{62}$ of the mini-batch D' **do**
Compute DQN's average loss $L_{t'}(\theta) = \frac{1}{2}\Big(r(t') + \max_{\mathsf{m}^*} Q^t\big(\mathsf{c}(t'+1),\mathsf{m}^*,\theta^t\big) - \max_{\mathsf{m}^*} Q\big(\mathsf{c}(t'),\mathsf{m}^*,\theta\big)\Big)^2$,
Perform a gradient descent step on $L_{t'}(\theta)$ w.r.t θ
end
Update the target network $Q^t(\cdot)$ every 1000 steps
end

In 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:

$$\min_{\mathcal{P}_i} \quad \sum\nolimits_{j=1}^{B}\sum\nolimits_{\mathsf{o}_i(t)\in\mathcal{P}_{i,j}}\Big|\pi_i^*(\mathsf{o}_i(t))-\mu_{i,j}\Big|, \tag{5}$$

where $\mathcal{P}_i=\{\mathcal{P}_{i,1},...,\mathcal{P}_{i,B}\}$ is a partition of Ω, and

$$\pi_i^*(\mathsf{o}_i(t)) = \text{argmax}_{\mathbf{m}_i^*}\ p_{\pi^*}(\mathbf{m}_i^*|\mathsf{o}_i(t)), \tag{6}$$

and $\mathbf{m}_i^*$ is the optimal action of agent i, which is i-th element of $\mathbf{m}^*\triangleq\pi^*\big(\mathbf{o}_1(t),...,\mathbf{o}_N(t)\big)$. Thus $\pi_i^*(\mathsf{o}_i(t))$ is the maximum aposteriori estimator of $\mathbf{m}_i^*=\pi^*\big(\mathbf{s}(t)\big)$ given the local observation $\mathsf{o}_i(t)$.

Once the clustering in (5) 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 $\forall k\in\{1,...,B\}$: $\pi_i^{c,ABSA-2}(\mathbf{o}_i)=\mathbf{c}^{(k)}$ 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 deep Q-network (DQN) architecture at the CC.

IV. Performance Evaluation

In this section, we evaluate our proposed schemes via numerical results for a special case of geometric consensus problems [13] - the multi-agent rendezvous problem [6] [5], in which the communication channel between the agents and the CC has a limited bit-budget. At each time t, given each agent i's observation $\mathsf{o}_i(t)$, all agents receive a single team reward

$$\mathrm{r}_t=\begin{cases} C_1, & \text{if } \exists\, i,j\in N:\mathsf{o}_i(t)\in\Omega^T\ \ \&\ \ \mathsf{o}_j(t)\notin\Omega^T\\ C_2, & \text{if } \nexists\, i\in N:\mathsf{o}_i(t)\in\Omega-\Omega^T,\\ 0, & \text{otherwise,}\end{cases} \tag{7}$$

where $C_1<C_2$ and Ω^T is the set of terminal observations - i.e., the episode terminates if $\exists\, i\in\mathcal{N}:\mathsf{o}_i(t)\in\Omega^T$. 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 [6].

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 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 Adam optimizer. An experience reply 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 ϵ-greedy with the initial $\epsilon=0.05$ and final $\epsilon=0.005$ is used [14]. 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.

For all black curves, one prior centralized training phase to obtain $\pi^*(\cdot)$ is required. As detailed in Section III, ABSA leverages $\pi^*(\cdot)$ to design π^c and then π^m afterwards. Dashed curves, HOC and HNC, as proposed by [6] provide heuristic schemes which exploit the domain knowledge of its designer

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

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 and HOC require communication/coordination between agents prior to the starting point of the task - which is not required for any other scheme. To obtain the results demonstrated in Fig. 4, 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 π^m after π^c is already computed using equations (4) and (5). 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 observation process, ABSA-2 was proposed which is seen to approach to the optimal solution as d grows. All ABSA-2 curves are plotted with $B = 3$, and ABSA-1 curve is plotted with $B = |\mathcal{M}|^N = 25$ in the two agent scenario - Fig. 2 - and $B = |\mathcal{M}|^N = 125$ in 3 agent scenarios - Fig. 3 and 4.

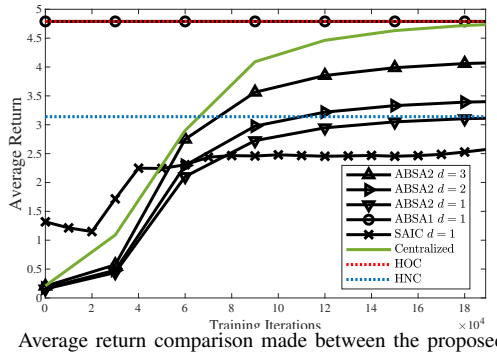

Figure 4. Average return comparison made between the proposed schemes and some benchmarks introduced in [6] - three agent scenario under a constant bit-budget values.

In Fig. 5, we see how the performance of ABSA-2 compares with HNC, HOC and SAIC at different rates of quantization. As expected, with the increase in the number of communication symbols, 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 in the lowest value of bit-budget where SAIC's performance is dropped drastically. 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 increased computational complexity of ABSA-2.

V. Conclusion

In this paper, we have investigated the joint design of the control and communication policies in multi-agent system under centralized control and distributed communication policies. We first proposed an action-based state aggregation algorithm (ABSA-1) for lossless compression with optimal average return performance. 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

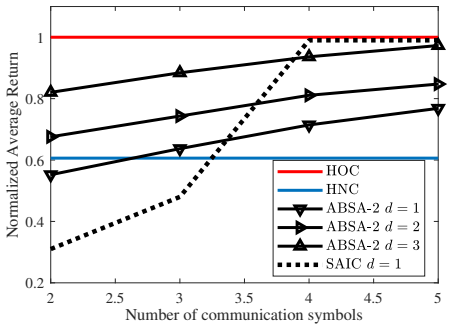

Figure 5. Normalized return comparison made between the proposed schemes and some benchmarks introduced in [6] - two agent scenario under four different bit-budget $|\mathcal{C}|$ values.

the proposed algorithms via numerical experiments performed on a geometric consensus via a number of representative metrics. Furthermore, our numerical studies demonstrates 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 the future research.

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] 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 GLOBECOM*, 2021, pp. 1–6.

[3] 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.

[4] A. Mostaani, T. X. Vu, S. K. Sharma, V.-D. Nguyen, Q. Liao, and S. Chatzinotas, "Task-oriented communication design in cyber-physical systems: A survey on theory and applications," *IEEE Access*, vol. 10, pp. 133 842–133 868, 2022.

[5] C. E. Shannon and W. Weaver, "The mathematical theory of communication [1949]. urbana, il," 1959.

[6] 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.

[7] 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 JSAC*, vol. 39, no. 8, pp. 2590–2603, 2021.

[8] 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.

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

[10] A. Mostaani, O. Simeone, S. Chatzinotas, and B. Ottersten, "Learning-based physical layer communications for multiagent collaboration," in *2019 IEEE Intl. Symp. PIMRC*, Sep. 2019.

[11] 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.

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

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

[14] V. Mnih, K. Kavukcuoglu, and et al, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.