Context-Aware Effective Communications

Tze-Yang Tung, Szymon Kobus, Deniz Gündüz
Information Processing and Communications Laboratory (IPC-Lab)
Dept. of Electrical and Electronic Engineering, Imperial College London, UK
{tze-yang.tung14, szymon.kobus17, d.gunduz}@imperial.ac.uk

***Abstract*—We investigate the *effective communications* problem, in which the goal of the transmitter is to impact the actions of the receiver over a time frame in order to maximize the prescribed reward function. This problem is formulated as a multi-agent partially-observable Markov decision process (MA-POMDP), where one agent can communicate to the other through a noisy communication channel. In this paper, we show that not only should the communication scheme be jointly designed with the underlying learning objective, but the *context* of the problem can also be exploited to achieve even greater effectiveness. Here, the *context* refers to a function of the state of the environment that is available to both agents. We then show that, using different communication schemes depending on the context is beneficial to the effectiveness of the solution. We emphasize that this is different from sending different messages at different states; with the proposed context-aware communication scheme, the same message is interpreted differently by the receiver depending on the context, similarly to human communications, where the meaning of a sentence may change depending on the context.**

***Index Terms*—Reinforcement learning, learn to communicate, semantic communications, effective communications, multi-agent systems, deep learning.**

I. Introduction

In his article "Recent Contributions to the Mathematical Theory of Communication", which appeared in the same volume with Shannon's "The Mathematical Theory of Communications" in [1], Weaver categorized communication problems into three levels:

1) *Technical problem*: How accurately can the symbols of communication be transmitted?
2) *Semantic problem*: How precisely do the transmitted symbols convey the desired meaning?
3) *Effectiveness problem*: How effectively does the received meaning affect conduct in the desired way?

Following Shannon's original formulation, communication systems have been designed with the *technical problem* as its sole focus. This allows the separation of the high level problem, such as the control signals to be sent in a swarm of drones, from the technical communication problem, such as the delivery of these signals to the desired recipient [2]. Shannon's theorem [1], which proved that the separation of source and channel coding is without loss of optimality when the blocklength goes to infinity, can be seen as the theoretical basis for the technical communication problem should be treated exclusively. However, this approach can lead to inefficiencies and delays, making certain control problems that require ultra-low latency, such as autonomous vehicle-to-vehicle coordination [3], challenging.

There is an increasing interest for future communication systems to adopt a goal-oriented approach [4], [5]. The idea is for the technical communication problem to be solved jointly with the high level problem, such that the communication allows the goal of the communication to be achieved most efficiently. In many cases, the goal can be considered as a general fidelity metric imposed on the reconstruction at the receiver. For example, in many learning applications, the goal of the receiver is to carry out some inference on the source signal, rather than reconstructing it [6]. In these problems, which can be considered within the *semantic communication* framework, the inferred parameter can be considered as the underlying source signal to be reconstructed, and the overall problem can be formulated as a remote joint source-channel coding problem. In this paper, we instead consider the *effectiveness problem* that goes beyond reconstructing a latent variable that may or may not be observable to the encoder.

An early work by Goldreich et al. [7], put forward a general theory of goal-oriented communication, where they defined "reliable communication" as means to overcome any "misunderstanding" between parties towards achieving a given goal with the help of signals received from the environment. This is closer to the *effectiveness problem* in Weaver's classification. Another formulation of the goal-oriented communication can be within multi-agent reinforcement learning (MARL) settings, where agents communicate with each other to accomplish a goal collaboratively [8]–[19]. Particularly, in [19], an effective communication framework is introduced, where communication enables agents to collaboratively solve a task, which can be formulated as a multi-agent partially observable Markov decision process (MA-POMDP). The authors showed that by jointly solving the high level MDP with the communication problem using a reinforcement learning framework, the agents are able to learn a communication scheme which achieves the goal of the MDP most effectively.

In this work, we show that not only does a communication scheme that most effectively solves a particular high level MDP problem arises from the joint optimization of communication and learning problems, the *context* of the problem can also be exploited, such that the communication scheme itself changes depending on the context. To demonstrate this, we consider a MA-POMDP problem, where the agents must communicate with each other over a noisy communication channel to collaboratively achieve a task. We then define the *context* of the problem as a function of the environment state, and provide a solution, in which changing the

communication scheme depending on the context improves the performance. Finally, we solve the MA-POMDP using reinforcement learning (RL) techniques, such that the agents learn a communication scheme that solves the POMDP and the communication problem jointly. Our main contributions are as follows:

1) For a particular MA-POMDP problem with one-way communication, we propose a heuristic solution, and show that changing the communication scheme depending on the context can improve the end performance.
2) Using RL, we obtain solutions that are jointly solved with the higher level POMDP, and show that they outperform the heuristic solution.

II. Related Works

Goal oriented communication has received increasing interest recently [5]. The requirement for low latency and ultra reliable communications in settings such as factory automation [20], autonomous vehicle-to-vehicle communications [3], [21], and remote image classification [6], has given rise to the demand for future generations of communication systems to be jointly designed with the problems they intend to solve. An early work by Goldreich et al. [7], developed a framework that redefined reliable communication as overcoming any "misunderstanding" between parties towards achieving a given goal, rather than reproducing "at one point either exactly or approximately a message selected at another point", as *technical* solutions to communications has been designed for.

In parallel, there are growing number of works studying the emergence of languages among agents in a multi-agent system that can arise from the need to coordinate and collaborate to accomplish a goal. Foerster et al. [8] investigated the role of communications in cooperative multi-agent systems and the type of machine learned "languages" that may arise from it. Many subsequent works have used communications as a means to coordinate policies of individual agents [9]–[18]. However, these works generally treat the communication channel as a perfect bit pipe, ignoring channel imperfections.

The *effective communication* problem over a noisy channel is formulated in [19], [22]. Here, unlike in [3], [5], [6], [20], [21], goal-oriented communication is defined over a time-horizon, where communication takes place in steps, and the goal in each step is to maximize the accumulated reward over the time horizon. In this framework, two or multiple agents must communicate via a noisy communication channel in order to coordinate and accomplish a goal. As such, the communication scheme must be learned jointly with the MARL policy. A similar problem was also considered in [23] in the context of emergent communications among agents communicating over noisy channels. It is shown in [19], [22] that by considering the communication scheme and the learning algorithm jointly, the resultant policy is more robust to channel errors than a policy that combines the learning algorithm with an existing communication scheme. However, they do not exploit the context of the problem as the solutions to the problems introduced in those works can be deconstructed to a sequence of instructions to the other agent. It can be argued that the joint learning and communication scheme that arise from [19], [22] are basically joint source-channel coding (JSCC) schemes that are more robust to channel distortions, with the difference that the quality measure in defining the JSCC scheme depends on the value function derived from the underlying POMDP, rather than being specified externally.

In the context-aware communications literature, there have been some works that utilize contextual information in vehicular to infrastructure (V2X) networks to improve quality of service (QoS), such as traffic routing and road safety [24]. In [25], contextual information, such as the number of vehicles on a road, is used to improve communication network congestion in a V2X network. However, the contextual information is used only to update the parameters in a communication scheme, while the scheme itself remains the same. Our goal in this work is to show that the contextual information, which we will define in the sequel, in a MA-POMDP can be utilized such that the communication scheme changes depending on the context in order to achieve the goal most effectively.

III. Problem Formulation

Consider a MA-POMDP with two agents, defined by $(\mathcal{S}, \{\mathcal{O}_i\}_{i=1}^2, \{\mathcal{A}_i\}_{i=1}^2, P, r)$, where $\mathcal{S}$ represents all possible states of the environment, $\mathcal{O}_i$ and $\mathcal{A}_i$ are the observation and action sets of agent $i = 1, 2$, respectively. P is the transition kernel that governs the environment, and $r : \mathcal{S} \times \prod_{i=1}^2 \mathcal{A}_i \mapsto \mathbb{R}$ is the reward function. Although in general the reward function for each agent can be different, herein we consider a fully cooperative environment with a common reward function. To coordinate, the agents are endowed with a noisy communication channel. Let the channel be defined by a conditional probability distribution $P_c(\hat{\mathbf{m}}^{(t)}|\mathbf{m}^{(t)})$, where $\mathbf{m}^{(t)} = (m_1^{(t)}, m_2^{(t)})$ are the transmitted messages from agent 1 and 2, respectively, at time step t, and $\hat{\mathbf{m}}^{(t)} = (\hat{m}_1^{(t)}, \hat{m}_2^{(t)})$ are the received messages. This communication channel is independent of the environment, such that, the environment transitions based only on environmental actions, and the only impact of the communication channel is that the actions of the agents can now depend on the received messages as well. We define the received message as part of the observation and the transmitted message as part of the action taken. As such, at each time step t, agent i makes an observation $\mathbf{o}_i^{(t)} = (e_i^{(t)}, \hat{m}_i^{(t)}) \in \mathcal{O}_i$, where $e_i^{(t)} \in \mathbf{s}^{(t)}$ is the environmental observation, and takes an action $\mathbf{a}_i^{(t)} = (a_i^{(t)}, m_i^{(t)}) \in \mathcal{A}_i$, where $a_i^{(t)}$ is the environmental action taken by agent i. The state of the MA-POMDP then transitions from $\mathbf{s}^{(t)}$ to $\mathbf{s}^{(t+1)}$ based on the joint actions of the agents via the probability transition kernel $P(\mathbf{s}^{(t+1)}|\mathbf{s}^{(t)}, \mathbf{a}_1^{(t)}, \mathbf{a}_2^{(t)})$. The observations in the next time step follow the conditional distribution $P(\mathbf{o}^{(t+1)}|\mathbf{s}^{(t)}, \mathbf{a}_1^{(t)}, \mathbf{a}_2^{(t)})$. Each agent then receives a common reward based on the state and actions $r(\mathbf{s}^{(t)}, \mathbf{a}_1^{(t)}, \mathbf{a}_2^{(t)})$.

In this paper, to simplify the presentation, we will consider a particular 2D world example of size $H \times W$. Inside the world, there is a scout agent ($i = 2$) and a treasure, which

is located on one of the integer grid points in $\mathcal{G} = [H] \times [W]$, where $[G] = \{0, \dots, G\}$. As such, the treasure can only exist on a finite number of locations, whereas the world itself is a 2D continuous plane. The scout knows its own location in the world but does not know the location of the treasure. A guide agent ($i = 1$), which can observe the state of the world must communicate with the scout agent in order for the scout agent to arrive at the treasure as quickly as possible. Let the position of the scout agent at time step t be denoted by $\mathbf{p}^{(t)} = (p_x^{(t)}, p_y^{(t)}) \in \mathbb{R}^{H \times W}$ and that of the treasure by $\mathbf{g} = (g_x, g_y) \in \mathcal{G}$. As such, the state of the environment at time t is defined as $\mathbf{e}^{(t)} = (\mathbf{p}^{(t)}, \mathbf{g})$. We define the action set of the guide to be a codebook $\mathcal{C} \in \mathbb{C}$ (i.e., $\mathcal{A}_1 = \mathcal{C}$) and the action set of the scout to be $\mathcal{A}_2 = [-k, k]^2$, meaning it can take a maximum step size of k in both the x and y directions.

At each time step t, the guide makes an observation $\mathbf{o}_1^{(t)} = \mathbf{e}^{(t)}$ and chooses a message $m_1^{(t)}$ from the codebook $\mathcal{C} \in \mathbb{C}$, as its action based on its policy $m_1^{(t)} = \pi_1(\mathbf{o}_1^{(t)})$, where $\pi_1 : \mathcal{O}_1 \mapsto \mathcal{C}$. The message is transmitted to the scout via an additive white Gaussian noise (AWGN) channel, $\hat{m}_2^{(t)} = m_1^{(t)} + z^{(t)}$, where $z^{(t)} \sim CN(0, \sigma_z^2 \mathbf{I})$ is complex Gaussian distributed with zero mean and covariance $\sigma_z^2 \mathbf{I}$. We impose a per channel use power constraint on the message to be $|m_1^{(t)}|^2 \leq 1$. The received message forms part of the observation of the scout $\mathbf{o}_2^{(t)} = (\hat{m}_2^{(t)}, \mathbf{p}^{(t)}, \xi^{(t)})$, where $\xi^{(t)} = \mathbb{I}_{\{||\mathbf{g} - \mathbf{p}^{(t)}||_\infty \leq k/2\}}$ is an indicator for when the scout is within $k \times k$ square around the treasure. The scout then takes an action $\mathbf{a}_2^{(t)} = (a_x^{(t)}, a_y^{(t)}) \in \mathcal{A}_2$ based on its policy such that $\mathbf{a}_2^{(t)} = \pi_2(\mathbf{o}_2^{(t)})$, where $\pi_2 : \mathcal{O}_2 \mapsto \mathcal{A}_2$, and updates its position to $\mathbf{p}^{(t+1)} = (p_x^{(t)} + a_x^{(t)}, p_y^{(t)} + a_y^{(t)})$. The episode terminates when $||\mathbf{g} - \mathbf{p}^{(t)}||_\infty \leq 0.5$; that is, the scout is within the 0.5×0.5 unit square around the treasure.

Given the above MA-POMDP, we define the state as $\mathbf{s}^{(t)} = (\mathbf{g}, \mathbf{p}^{(t)}, \hat{m}_2^{(t)}, \xi^{(t)})$. Also, given the power constraint of the message $|m_1^{(t)}|^2 \leq 1$, we define the signal-to-noise ratio (SNR) of the communication channel as $\text{SNR} = 10 \log_{10} \left(1/\sigma_z^2\right)$ dB.

We measure the effectiveness of the agents' policy by considering the probability of successfully arriving at the treasure within a certain number of time steps. This is analogous to the block error rate (BLER) for a particular channel codeword length in the *technical* level, except here we are concerned with the *effectiveness* of the policy with respect to the "goal". The reward is therefore defined to encourage the fewest number of steps to reach the treasure and will be specified in the sequel.

We can define the *value function* $V_\Pi(\mathbf{s})$ and the *state-action function*, also known as the *Q-function*, $Q_\Pi(\mathbf{s}, \mathbf{a})$ as

$$V_\Pi(\mathbf{s}) = \mathbb{E}_\Pi \left[\sum_{t=1}^{\infty} \gamma^{t-1} r^{(t)} \middle| \mathbf{s}^{(1)} = \mathbf{s} \right], \tag{1}$$

$$Q_\Pi(\mathbf{s}^{(t)}, \mathbf{a}^{(t)}) = \mathbb{E}_\Pi \left[\sum_{n=t}^{\infty} \gamma^{(n-t)} r^{(t)} \middle| \mathbf{s}^{(t)}, \mathbf{a}^{(t)} \right], \tag{2}$$

where $\mathbf{a}^{(t)} = (\mathbf{a}_1^{(t)}, \mathbf{a}_2^{(t)})$ and $\Pi = \pi_1 \times \pi_2$ is the joint policy of the 2 agents. The expected return from the distribution of initial states can then be defined as

$$J(\Pi) = \mathbb{E}_{\mathbf{s}^{(t)} \sim \rho^\Pi, \mathbf{a}^{(t)} \sim \Pi} \left[\sum_{t=1}^{\infty} \gamma^{t-1} r^{(t)}(\mathbf{s}^{(t)}, \mathbf{a}^{(t)}) \right], \tag{3}$$

where ρ^Π is the discounted state visitation distribution for the joint policy Π. The optimization problem can be defined as

$$\Pi^* = \arg\max_\Pi J(\Pi) \tag{4}$$

IV. Proposed Solutions

A. Heuristic Solution

Herein, we construct a two-phase solution, where a different communication scheme is employed in each phase, and the agents switch between the two phases depending on the context of the MA-POMDP, as shown in Algorithm 1. The notation $\text{diag}(\mathbf{A})$ refers to the main diagonal of a square matrix $\mathbf{A}$, $\text{sgn}(x)$ refers to the sign of the value x, and $\text{clip}(x, a, b)$ clips the value x between a and b, for $a < b$. We endow each agent with memory, denoted by $\mathbf{b}_1^{(t)}$ and $\mathbf{b}_2^{(t)}$, with $\mathbf{b}_2^{(t)} = E[\mathbf{g}|\hat{\mathbf{g}}^{(t)}, \dots, \hat{\mathbf{g}}^{(0)}]$ the minimum mean-squared error (MMSE) estimate of the treasure location based on all prior estimates $(\hat{\mathbf{g}}^{(t)}, \dots, \hat{\mathbf{g}}^{(0)})$, where $\hat{\mathbf{g}}^{(t)}$ is the minimum variance unbiased (MVU) estimate of the treasure location based only on the message from time step t. We will refer to $\mathbf{b}_2^{(t)}$ as the scout's *belief* on the location of the treasure at time step t. Then $\mathbf{b}_1^{(t)} = E[\text{diag}((\mathbf{g} - \mathbf{b}_2^{(t)})(\mathbf{g} - \mathbf{b}_2^{(t)})^\top)|\mathbf{p}^{(t)}, \dots, \mathbf{p}^{(0)}]$ is the mean-squared error (MSE) of the scout's belief. At each time step t, the guide transmits the difference between the treasure location $\mathbf{g}$ and the scout's location $\mathbf{p}^{(t-1)}$, denoted by $\mathbf{u}^{(t)}$, to the scout. A message $v^{(t)}$ is first formed by pairing the x and y values of $\mathbf{u}^{(t)}$ to form a complex symbol, such that $v^{(t)} = u_x^{(t)} + j u_y^{(t)}$. To satisfy the average power constraint $E[m_1^{(t)}] \leq 1$, the transmitted message $m_1^{(t)}$ is normalized by scaling it as $m_1^{(t)} = d^{(t)} v^{(t)}$, where $d^{(t)} = 1/\sigma^{(t)}$ and $\sigma^{(t)} = E[||\mathbf{u}^{(t)}||_2 | \mathbf{p}^{(t-1)}, \mathbf{b}_1^{(t-1)}]$ is the expected magnitude of $\mathbf{u}$ given the scout's location in the previous time step and the MSE of the scout's belief. The scout then updates its belief based on the new MVU estimate of the treasure location $\hat{\mathbf{g}}^{(t)}$ from the message received in this time step $\hat{m}_2^{(t)}$. Note that here we assume the scout knows the scaling factor $d^{(t)}$ for computing its belief $\mathbf{b}_2^{(t)}$.

It can be seen from Algorithm 1 that there are two phases of the solution, separated by conditions involving the expected l_2 distance to treasure $\sigma^{(t)}$ and the scout's l_∞ distance to the treasure $||\mathbf{u}^{(t)}||_\infty$. As such, we can view the *context* in each step as a function of the joint history of observations $\{(\mathbf{o}_1^{(j)}, \mathbf{o}_2^{(j)})\}_{j=1}^t$. The key difference between the two phases is that, in the first phase, the scout computes the MMSE estimate of the location of the treasure $\mathbf{g}$ to determine the best action to take, whereas in the second phase, it switches to the maximum likelihood estimate (MLE). In both phases, the scout takes the action that brings it closest to its estimate of the treasure location. An intuitive explanation for why this two-phased scheme is better than using only the first phase is that, initially, the scout is unsure about where the treasure

Algorithm 1: Heuristic context-aware communication scheme.

Initialize:
$\mathbf{p}^{(0)} = (H/2, W/2)$, $\mathbf{b}_1^{(0)} = \frac{(2G+1)^2-1}{12}$, $\xi^{(t)} = 0$, $t = 0$
Start phase-1:
while $\sigma^{(t)} \geq (4\sqrt{2}\sigma_z)^{-1}$ *and* $\xi^{(t)} = 0$ **do**
- $t = t+1$
- $\mathbf{u}^{(t)} = (u_x^{(t)}, u_y^{(t)}) = \mathbf{g} - \mathbf{p}^{(t-1)}$
- $\sigma^{(t)} = E[||\mathbf{u}^{(t)}||_2 | \mathbf{p}^{(t-1)}, \mathbf{b}_1^{(t-1)}]$
- $d^{(t)} = \frac{1}{\sigma^{(t)}}$
- $v^{(t)} = u_x^{(t)} + j u_y^{(t)}$
- $m_1^{(t)} = d^{(t)} v^{(t)}$
- $\hat{m}_2^{(t)} = m_1^{(t)} + z^{(t)}$
- $\hat{v}^{(t)} = \hat{m}_2^{(t)} / d^{(t)}$
- $\hat{\mathbf{u}}^{(t)} = (\text{Re}\{\hat{v}^{(t)}\}, \text{Im}\{\hat{v}^{(t)}\})$
- $\hat{\mathbf{g}}^{(t)} = \mathbf{p}^{(t-1)} + \hat{\mathbf{u}}^{(t)}$
- $\mathbf{b}_2^{(t)} = E[\mathbf{g} | \hat{\mathbf{g}}^{(t)}, \ldots, \hat{\mathbf{g}}^{(0)}]$
- $\mathbf{b}_1^{(t)} = E[\text{diag}((\mathbf{g} - \mathbf{b}_2^{(t)})(\mathbf{g} - \mathbf{b}_2^{(t)})^\top) | \mathbf{p}^{(t)}, \ldots, \mathbf{p}^{(0)}]$
- $\mathbf{a}_2^{(t)} = \text{clip}(\mathbf{b}_2^{(t)} - \mathbf{p}^{(t-1)}, k, -k)$
- $\mathbf{p}^{(t)} = \mathbf{p}^{(t-1)} + \mathbf{a}_2^{(t)}$
- $\xi^{(t)} = \mathbb{I}_{\{||\mathbf{g} - \mathbf{p}^{(t)}||_\infty \leq k/2\}}$

Start phase-2:
while $||\mathbf{g} - \mathbf{p}^{(t)}||_\infty > 0.5$ *and* $\xi^{(t)} = 1$ **do**
- $t = t+1$
- $\mathbf{u}^{(t)} = \mathbf{g} - \mathbf{p}^{(t-1)}$
- $d^{(t)} = 2\exp\left(\frac{(d^{(t-1)})^2}{16\sigma_z^2}\right)$
- $v^{(t)} = u_x^{(t)} + j u_y^{(t)}$
- $m_1^{(t)} = d^{(t)} v^{(t)}$
- $\hat{m}_2^{(t)} = m_1^{(t)} + z^{(t)}$
- $\hat{v}^{(t)} = \hat{m}_2^{(t)} / d^{(t)}$
- $\hat{\mathbf{u}}^{(t)} = (\text{Re}\{\hat{v}^{(t)}\}, \text{Im}\{\hat{v}^{(t)}\})$
- $\hat{\mathbf{g}}^{(t)} = \mathbf{p}^{(t-1)} + \hat{\mathbf{u}}^{(t)}$
- $\mathbf{b}_2^{(t)} = E[\mathbf{g} | \hat{\mathbf{g}}^{(t)}, \ldots, \hat{\mathbf{g}}^{(0)}]$
- $\mathbf{p}^{(t)} = \arg\max_{\mathbf{g} \in \mathcal{G}} P(\mathbf{b}_2^{(t)} | \mathbf{g})$

is, characterized by the large MSE of its belief $\mathbf{b}_2^{(t)}$. As the MSE of the scout's belief decreases, the scout becomes more confident about the location of the treasure and it is better for it to choose the action that corresponds to the MLE. Indeed, we will show that switching between these two options will be beneficial.

Note that, in the grid world example considered in [19], the scout did not know either its own position, or that of the treasure. Therefore, the only option for the guide was to send the actions to be taken by the scout. However, when the scout knows its own location, the guide can also send the treasure location (approximately) for the scout to move towards it, as can be seen by the solution provided herein.

To prove that the two-phased scheme is better than a scheme using only the first phase, we first introduce a lemma.

Lemma 1: [26] For any $d \geq 4\sigma_z$, let U be a uniform d-quantization of a Gaussian random variable $Z \sim N(0, \sigma_z^2)$, in the sense that for each integer l, if $Z \in (dl - d/2, dl + d/2]$, then $U = l$. Then, $E[U^2]$ is upper-bounded by

$$E[U^2] \leq \frac{1.6\sigma_z}{d} \exp\left(\frac{-d^2}{8\sigma_z^2}\right). \tag{5}$$

Given Lemma 1 and Algorithm 1, at time step $t = n$, if $d^{(n)} \geq 4\sqrt{2}\sigma_z$ and the distance to the treasure $||\mathbf{u}^{(n)}||_\infty \leq k/2$, then, with a slight exception, the scout's position based on the MLE $\mathbf{p}^{(n)} = \arg\max_{\mathbf{g} \in \mathcal{G}} P(\mathbf{b}_2^{(n)} | \mathbf{g})$ is equivalent to a $d^{(n)}$-quantization of the belief $\mathbf{b}_2^{(n)}$ in 2D. The exception is due to the finite set of goal locations $\mathcal{G}$, but since this restriction only serves to reduce the second moment, *Lemma 1* still holds. Moreover, Algorithm 1 also implies the following lemma:

Lemma 2: [26] If $d^{(n)} \geq 4\sqrt{2}\sigma_z$ and $\xi^{(n)} = 1$, for $t \geq n$, Algorithm 1 satisfies:

$$\sum_{t=n}^{\infty} E[|m_1^{(t)}|^2] \leq 5, \text{ and} \tag{6}$$

$$P(\mathbf{p}^{(t)} \neq \mathbf{g}) \leq \frac{1}{h_{t-n+1}(2\sigma_z^2)}, \tag{7}$$

where $h_t(x) = \exp(\cdots(\exp(x))\cdots)$ with t exponentials.

Therefore, not only the total power of the messages transmitted in the second phase of Algorithm 1 is finite, and therefore, satisfy the average power constraint, but also the probability that the scout does not arrive at the treasure decreases in the order of $t - n + 1$ exponentials. This is in contrast to the first order exponential decrease in probability of the scout not arriving at the treasure if the scout had continued to use the MMSE strategy of phase-1:

$$P_{\text{MMSE}}(||\mathbf{g} - \mathbf{p}^{(t)}||_\infty \leq 0.5) \leq 2Q(\boldsymbol{\gamma}_t), \tag{8}$$

where $\boldsymbol{\gamma}_t = \frac{1}{2\sqrt{\mathbf{b}_1^{(t)}}}$,

$$Q(\mathbf{x}) = \frac{1}{2\pi} \int_{u_1 = x_1}^{\infty} \int_{u_2 = x_2}^{\infty} e^{\frac{-(u_1^2 + u_2^2)}{2}} du_1 du_2,$$

and $\mathbf{x} = (x_1, x_2)$. Therefore, switching from phase-1 to phase-2 of Algorithm 1 leads to the scout finding the treasure faster on average. Note that, this result is based on the average power constraint $E[|m_1^{(t)}|^2] \leq 1$ being met, whereas the problem formulation in Sec. III calls for a per channel use power constraint $|m_1^{(t)}|^2 \leq 1$. To meet the per channel use power constraint, we scale the gain $d^{(t)}$ such that $|m_1^{(t)}|^2 \leq 1$. Although Lemma 2 no longer holds with this constraint, we observe in our numerical results in Sec. V that phase 2 is still beneficial.

B. Reinforcement Learning

We now detail the RL method used to optimize Eqn. (4). Since the objective is to reach the goal in the least number of steps, we define the reward function as

$$r^{(t)} = r_{end}^{(t)} + r_{dist}^{(t)}, \tag{9}$$

where

$$r_{end}^{(t)} = \begin{cases} 10, & \text{if } ||\mathbf{g} - \mathbf{p}^{(t)}||_\infty \leq 0.5, \\ 0, & \text{otherwise,} \end{cases} \tag{10}$$

which encourages the scout to reach the treasure in the least number of steps, and

$$r_{dist}^{(t)} = -\frac{10}{\text{steps}_{\max}} \log(||\mathbf{u}^{(t)}||_1 + 1), \tag{11}$$

learning rate	10^{-3}
γ	0.99
τ	0.05
batch size	1024
exploration noise decay ν	1500
nn layer shape	[64,64,64(GRU cell)]
nn hidden activation function	leaky ReLU
nn out activation function	tanh
$\text{steps}_{\max}$	$[40, 45, 50, 55]$

TABLE I
RL TRAINING HYPERPARAMETERS.

which encourages the scout to move towards the treasure in each step. Here, $\text{steps}_{\max}$ is the maximum number of steps per episode. The logarithm is introduced to normalize the magnitude of the reward.

The policy of each agent is implemented as a recurrent neural network (RNN), with hidden states $\mathbf{b}_i^{(t)}$ for $i = 1, 2$, such that the hidden states act as the memory of the policy, as in the heuristic solution shown in the previous section. To deal with non-stationarity when learning with multiple agents in an environment, the agents are optimized jointly. We employ an adapted recurrent multi-agent deep deterministic policy gradient (R-MADDPG) algorithm, proposed in [12]. Each agent is defined as an actor $\pi_{\boldsymbol{\theta}_i}$ network, parameterized by $\boldsymbol{\theta}_i$. We also utilize an RNN critic network $Q_{\boldsymbol{\phi}}$, parameterized by $\boldsymbol{\phi}$, to provide gradients for the update of the actor networks, with $\mathbf{h}^{(t)}$ being its hidden state.

During training, each agent keeps a reply buffer $\mathcal{B}_i$, containing experiences $(\mathbf{o}_i^{(t)}, \mathbf{o}_i^{(t+1)}, \mathbf{a}_i^{(t)}, \mathbf{a}_i^{(t+1)}, \mathbf{b}_i^{(t)}, \mathbf{b}_i^{(t+1)}, r^{(t)}, \mathbf{h}^{(t)}, \mathbf{h}^{(t+1)})$. The critic is updated via temporal difference (TD) learning loss

$$L(\boldsymbol{\phi}) = E_{\mathcal{B}_{1,2}}\Bigg[\Big(Q_{\boldsymbol{\phi}}\big(\{\mathbf{o}_i^{(t)}, \mathbf{a}_i^{(t)}\}_{i\in\{1,2\}}, \mathbf{h}^{(t)}\big) + r^{(t)} - \gamma Q_{\boldsymbol{\phi}^-}\big(\mathbf{o}_1^{(t+1)}, \bar{\mathbf{o}}_2^{(t+1)}, \bar{\mathbf{a}}_1^{(t+1)}, \bar{\mathbf{a}}_2^{(t+1)}, \mathbf{h}^{(t+1)}\big)\Big)^2\Bigg], \quad (12)$$

where

$$\bar{\mathbf{a}}_1^{(t+1)} = \pi_{\boldsymbol{\theta}_1^-}(\mathbf{o}_1^{(t+1)}, \mathbf{b}_1^{(t+1)}),\ \bar{\mathbf{a}}_2^{(t+1)} = \pi_{\boldsymbol{\theta}_2^-}(\bar{\mathbf{o}}_2^{(t+1)}, \mathbf{b}_2^{(t+1)}),$$

are actions chosen by the guide and scout, respectively, given the current policies. We define $\bar{\mathbf{o}}_2^{(t+1)} = (\bar{\mathbf{a}}_1^{(t+1)} + z, \mathbf{p}^{(t+1)}, \xi^{(t)})$, where $z \sim CN(0, \sigma_z^2\mathbf{I})$ is a complex Gaussian random variable, emulating the channel. The TD loss uses target networks, parameterized by $\boldsymbol{\phi}^-$ and $\boldsymbol{\theta}_i^-$, which are copies of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}_i$ but updated via a soft update $\boldsymbol{\theta}^- \leftarrow \tau\boldsymbol{\theta} + (1-\tau)\boldsymbol{\theta}^-$, where $0 \le \tau \le 1$, and $(\boldsymbol{\theta}, \boldsymbol{\theta}^-)$ are any set of network parameters and its target parameters. This soft update helps to stabilize the bootstrapped parameters in TD learning, as shown in Eqn. (12). To update the actor $\pi_{\boldsymbol{\theta}_i}$, the gradient is computed via the chain rule as shown in [27],

$$L(\boldsymbol{\theta}_{1,2}) = E_{\mathcal{B}_{1,2}}\Big[\nabla_{\boldsymbol{\theta}_{1,2}}\pi_{\boldsymbol{\theta}_2}((\pi_{\boldsymbol{\theta}_1}(\mathbf{o}_1^{(t)}, \mathbf{b}_1^{(t)}) + z, \mathbf{p}^{(t)}), \mathbf{b}_2^{(t)}) \nabla_{\mathbf{a}_{1,2}^{(t)}} Q_{\boldsymbol{\psi}}\big(\mathbf{o}_1^{(t)}, \bar{\mathbf{o}}_2^{(t)}, \bar{\mathbf{a}}_1^{(t)}, \bar{\mathbf{a}}_2^{(t)}, \mathbf{h}^{(t)}\big)\Big|_{\bar{\mathbf{a}}_i^{(t)} = \pi_{\boldsymbol{\theta}_i}(\mathbf{o}_i^{(t)}, \mathbf{b}_i^{(t)})}\Big]. \quad (13)$$

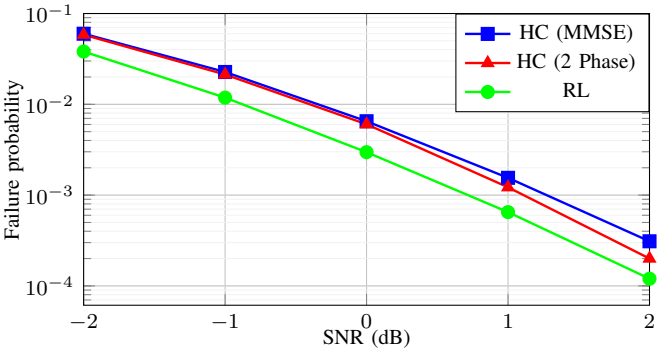

Fig. 1. Probability of the scout not arriving at the treasure within the episode ($\text{steps}_{\max} = 45$).

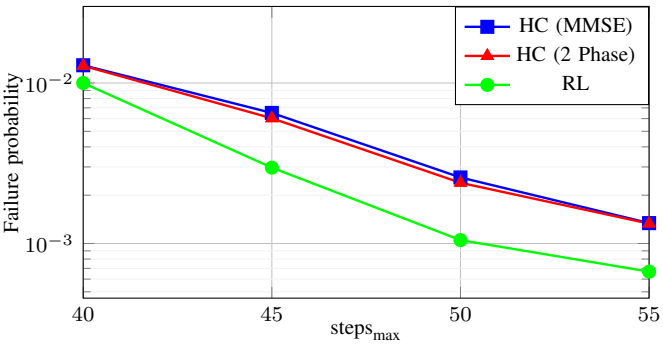

Fig. 2. Probability of the scout not arriving at the treasure within the episode for different $\text{steps}_{\max}$ (SNR = 0dB).

To promote exploration, we add noise to the actions taken during training as follows:

$$\mathbf{a}_i^{(t)} = \pi_{\boldsymbol{\theta}_i}(\mathbf{o}_i^{(t)}, \mathbf{b}_i^{(t)}) + \mathbf{w}^{(t)}, \quad (14)$$

where $\mathbf{w}^{(t)} \sim N(0, e^{-\frac{episode}{\nu}})$ is a Gaussian random variable with exponentially decaying variance with the episode number, and ν is a hyperparameter controlling the rate of decay.

As can be seen from Eqn. (12), the critic network $Q_{\boldsymbol{\phi}}$ utilizes full knowledge of the state of the MA-POMDP during training to avoid the non-stationarity of the environment from the perspective of any individual agent. During testing, the agents' policies $\pi_{\boldsymbol{\theta}_i}$ only utilize their respective observations $\mathbf{o}_i^{(t)}$ and memory states $\mathbf{b}_i^{(t)}$. Although the expectation is taken over the entire replay buffer in Eqns. (12) and (13), in practice, we uniformly sample mini-batches $b_i \sim \mathcal{B}_i$ and compute the empirical average of both losses.

V. NUMERICAL RESULTS

We consider a 512×512 world and initialize each episode by placing the scout at $\mathbf{p}^{(0)} = (H/2, W/2)$ and the treasure at a location uniformly randomly chosen on the integer grid of possible treasure locations $\mathbf{g} \sim U(\mathcal{G})$. We let the maximum step size of the scout be $k = 32$. The hyperparameters used to train the RL agents are shown in Table I. To observe the performance of each scheme with respect to a fixed number of channel uses, we limit the maximum number of steps per episode to $\text{steps}_{\max} \in [40, 45, 50, 55]$. We compare the performance of the RL solution with the heuristic solution in

Algorithm 1, denoted by "HC (2 Phase)", and the version that only uses the first phase with MMSE estimation, denoted by "HC (MMSE)". The results are obtained by averaging over one million episodes. Note that the "HC (MMSE)" scheme does not utilize the indicator $\xi^{(t)}$ that signifies that the scout is near the treasure. Moreover, when trained without the indicator $\xi^{(t)}$, the RL solution failed to converge, indicating that the context is important for the RL algorithm to solve the MA-POMDP effectively. We emphasize that the HC schemes assume the gain $d^{(t)}$ is known at each time step and that the scout knows the grid of possible treasure locations $\mathcal{G}$, whereas the RL solution must learn these through exploration since we are using a model free RL algorithm.

Fig. 1 shows the probability of failure to arrive at the treasure within 45 steps for each scheme. It can be seen that the RL solution performs much better than both HC solutions. It can also be seen from Fig. 1, that the probability of failure for the HC (2 Phase) scheme is lower than the HC (MMSE) scheme, as the proof in Sec. IV-A suggests. This shows the RL solution performs more *effectively* than the HC schemes, driven by the reward function we defined in Eqn. (9). Similar conclusions can be drawn from Fig. 2, where the failure probability is shown for different $\text{steps}_{\text{max}}$ at SNR $= 0$dB. It can be seen that the RL solution performs much better than both HC solutions. Moreover, as can be seen from both Figs. 1 and 2, the improvement induced by the HC (2 Phase) scheme can still be observed despite the constraint $|m_1^{(t)}|^2 \leq 1$. This is likely due to the probability that $|m_1^{(t)}|^2 > 1$ is small in the second phase when using the gain $d^{(t)} = 2\exp\left((d^{(t-1)})^2/16\sigma_z^2\right)$, as the MSE of the scout's belief would be small at this point, meaning $E[|m_1^{(t)}|^2] \approx |m_1^{(t)}|^2$.

VI. Conclusions

We have investigated the *effectiveness* problem in communications by considering a MA-POMDP with two agents, a guide and scout, where the guide needs to help the scout to reach a treasure in a grid world. We showed through a heuristic scheme that changing the communication protocol depending on the context, which is defined as a function of the agents' observations which they can agree on, can help improve the performance. Finally, we solve the MA-POMDP problem using RL techniques and numerically show that the obtained RL solution is more reliable on average than the heuristic schemes.

References

[1] C. E. Shannon and W. Weaver, *The mathematical theory of communication*. University of Illinois Press, 1963.

[2] M. Campion, P. Ranganathan, and S. Faruque, "UAV swarm communication and control architectures: a review," *Journal of Unmanned Vehicle Systems*, Nov. 2018.

[3] T.-H. Wang, S. Manivasagam, M. Liang, B. Yang, W. Zeng, J. Tu, and R. Urtasun, "V2VNet: Vehicle-to-vehicle communication for joint perception and prediction," in *European Conf. on Computer Vision (ECCV)*, Aug. 2020.

[4] P. Popovski, O. Simeone, F. Boccardi, D. Gündüz, and O. Sahin, "Semantic-Effectiveness Filtering and Control for Post-5G Wireless Connectivity," *Journal of the Indian Institute of Science*, vol. 100, pp. 435–443, Apr. 2020.

[5] E. Calvanese Strinati and S. Barbarossa, "6G networks: Beyond Shannon towards semantic and goal-oriented communications," *Computer Networks*, vol. 190, p. 107930, May 2021.

[6] M. Jankowski, D. Gündüz, and K. Mikolajczyk, "Deep joint source-channel coding for wireless image retrieval," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5070–5074, May 2020.

[7] O. Goldreich, B. Juba, and M. Sudan, "A theory of goal-oriented communication," *Journal of the ACM*, vol. 59, pp. 8:1–8:65, May 2012.

[8] J. N. Foerster, Y. M. Assael, N. de Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *Conf. on Neural Info. Proc. Systems (NIPS)*, Dec. 2016.

[9] A. Das, T. Gervet, J. Romoff, D. Batra, D. Parikh, M. Rabbat, and J. Pineau, "TarMAC: Targeted multi-agent communication," in *2019 International Conference on Machine Learning (ICML)*, June 2019.

[10] S. Havrylov and I. Titov, "Emergence of language with multi-agent games: Learning to communicate with sequences of symbols," in *Conf. on Neural Info. Proc. Systems (NIPS)*, Dec. 2017.

[11] P. Peng, Y. Wen, Y. Yang, Q. Yuan, Z. Tang, H. Long, and J. Wang, "Multiagent bidirectionally-coordinated nets: Emergence of human-level coordination in learning to play StarCraft combat games," *arXiv:1703.10069 [cs]*, Sept. 2017. arXiv: 1703.10069.

[12] R. E. Wang, M. Everett, and J. P. How, "R-MADDPG for partially observable environments and limited communication," in *2019 International Conference on Machine Learning (ICML)*, June 2019.

[13] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Conf. on Neural Info. Proc. Systems (NIPS)*, Dec. 2017.

[14] J. Blumenkamp and A. Prorok, "The emergence of adversarial communication in multi-agent reinforcement learning," in *2020 Conference on Robotic Learning (CoRL)*, Nov. 2020.

[15] Y. J. Park, Y. J. Lee, and S. B. Kim, "Cooperative multi-agent reinforcement learning with approximate model learning," *IEEE Access*, vol. 8, pp. 125389–125400, 2020.

[16] C. Sun, M. Shen, and J. How, "Scaling up multiagent reinforcement learning for robotic systems: Learn an adaptive sparse communication graph," in *IEEE/RSJ Int'l Conf. on Intel. Robots and Sys.*, Oct. 2020.

[17] E. Pesce and G. Montana, "Learning multi-agent coordination through graph-driven communication," in *2021 International Conference on Autonomous Agents and MultiAgent Systems*, p. 964–973, May 2021.

[18] S. Gupta, R. Hazra, and A. Dukkipati, "Networked multi-agent reinforcement learning with emergent communication," in *Int'l Conf. on Aut. Agents and Multiagent Sys.*, May 2020.

[19] T.-Y. Tung, S. Kobus, J. R. Pujol, and D. Gunduz, "Effective communications: A joint learning and communication framework for multi-agent reinforcement learning over noisy channels," *IEEE Trans. on Selected Areas in Comms.*, vol. 39, pp. 2590–2603, Aug. 2021.

[20] S. A. Ashraf, I. Aktas, E. Eriksson, K. W. Helmersson, and J. Ansari, "Ultra-reliable and low-latency communication for wireless factory automation: From LTE to 5G," in *IEEE Int'l Conf. on Emerging Tech. and Factory Automation*, pp. 1–8, Sept. 2016.

[21] N. Vadivelu, M. Ren, J. Tu, J. Wang, and R. Urtasun, "Learning to communicate and correct pose errors," in *2020 Conference on Robotic Learning (CoRL)*, Nov. 2020.

[22] J. S. P. Roig and D. Gündüz, "Remote reinforcement learning over a noisy channel," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, pp. 1–6, Dec. 2020.

[23] A. Mostaani, O. Simeone, S. Chatzinotas, and B. Ottersten, "Learning-based physical layer communications for multiagent collaboration," in *IEEE PIMRC*, pp. 1–6, Sept. 2019.

[24] J. Wan, D. Zhang, S. Zhao, L. T. Yang, and J. Lloret, "Context-aware vehicular cyber-physical systems with cloud support: Architecture, challenges, and solutions," *IEEE Communications Magazine*, vol. 52, pp. 106–113, Aug. 2014.

[25] M. Sepulcre, J. Gozalvez, J. Härri, and H. Hartenstein, "Contextual communications congestion control for cooperative vehicular networks," *IEEE Trans. on Wireless Comms.*, vol. 10, pp. 385–389, Feb. 2011.

[26] R. G. Gallager and B. Nakiboğlu, "Variations on a theme by Schalkwijk and Kailath," *IEEE Trans. on Info. Theory*, vol. 56, pp. 6–17, Jan. 2010.

[27] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *Int'l Conf. on Learning Repr. (ICLR)*, Nov. 2016.