Hurts to Be Too Early: Benefits and Drawbacks of Communication in Multi-Agent Learning

Parinaz Naghizadeh*, Maria Gorlatova†, Andrew S. Lan‡, Mung Chiang*
*Purdue University, †Duke University, ‡University of Massachusetts, Amherst
Emails: {parinaz, chiang}@purdue.edu, maria.gorlatova@duke.edu, andrewlan@cs.umass.edu

***Abstract*—We study a multi-agent partially observable environment in which autonomous agents aim to coordinate their actions, while also learning the parameters of the unknown environment through repeated interactions. In particular, we focus on the role of communication in a multi-agent reinforcement learning problem. We consider a learning algorithm in which agents make decisions based on their own observations of the environment, as well as the observations of other agents, which are collected through communication between agents. We first identify two potential benefits of this type of information sharing when agents' observation quality is heterogeneous: (1) it can facilitate coordination among agents, and (2) it can enhance the learning of *all* participants, including the better informed agents. We show however that these benefits of communication depend in general on its timing, so that delayed information sharing may be preferred in certain scenarios.**

***Index Terms*—Multi-agent reinforcement learning, information sharing, cooperative games.**

I. Introduction

The study of decentralized decision making under uncertainty by multiple autonomous agents arises in a wide variety of applications, including in wireless and telecommunication networks (e.g., opportunistic spectrum access, dynamic resource allocation), management of the smart grid and electricity markets, the operation of cyber-physical systems, and in a variety of other physical, social, and economic networks [1]–[7]. In these scenarios, the outcomes experienced by each agent is affected not only by their own decisions, but also by the actions taken by (a subset) of other agents in the system.

An instance of practical interest in these problems is when agents are willing to collaborate in order to achieve a common goal; these are referred to as cooperative games. In these environments, the literature has identified communication or information sharing as a way to facilitate *coordination* among agents, so that they can collaborate on reaching a shared goal. Examples include the study of communication for multi-robot formation control [5], autonomous vehicle coordination [6], and control of microgrids [4]. These works identify optimal communication schemes, agents' strategies, and the benefits of information sharing, under the assumption of some a priori knowledge about the environment and/or its dynamics.

However, the rise of self-organizing multi-agent systems in fully unknown environments, such as those arising in edge computing applications, has introduced an additional challenge for multi-agent systems. In particular, even in the absence of coordination problems, agents face the additional challenge of *learning* to act in the a priori unknown environment. For instance, a fleet of drones deployed for monitoring climate change [8] or for anti-poaching efforts in a wildlife area [9], or a team of disaster relief robots [10], will not only need to coordinate with one another, but at the same time need to learn features of the unknown environment in which they are deployed. The problem of learning to act through repeated interactions with an unknown environment, in the presence of other agents, is the subject of the multi-agent reinforcement learning literature; see [11] for a survey.

In this paper, we study the problem of multi-agent reinforcement learning in cooperative environments, and *aim to analytically evaluate the effects of information sharing on both the coordination and learning of the agents*. We are particularly interested in the role of communication when agents have heterogeneous capabilities in assessing their shared environment. This is motivated by the possible heterogeneity in agents' platforms; for instance, an agent might have a less accurate perception of the environment due to having weaker sensors, energy constraints, or limited storage. Such heterogeneity would be the case in fog computing [12]–[14], for example, where powerful cloud services and resource-limited edge nodes cooperate to assess the environment.

Specifically, we consider a collaborative, binary, partially observable environment, in which two agents receive independent observations about the state of the world. We assume that one of the agents is better informed, i.e., it makes an accurate observation of the true state of the environment. We analyze how enabling the sharing of these independent observations of differing quality between agents affects their decision making and learning.

We consider learning through a multi-agent version of the REINFORCE algorithm [15], which is a special case of actor-critic algorithms [16], [17], by extending it to incorporate communication between agents. The idea of using inter-agent communication for better learning has also been recently studied empirically in [18]–[20]. These works have proposed deep reinforcement learning methods based on actor-critic algorithms for multi-agent learning problems, with either policy parameter sharing [18] or full experience sharing [19], [20], and evaluate the performance of the resulting learning algorithms through empirical analysis. Our work, which only

This work was supported in part by the Waterman Award - NSF 1759652, the Comcast Innovation Fund Research Award, NSF CSR-1812797 grant, and Defense Advanced Research Projects Agency (DARPA) under contract No. HR001117C0052 and No. HR001117C0048.

requires sharing of environmental observations, provides a formal analysis of the various aspects in which communication can benefit agents, and more importantly, identifies its potential *drawbacks*, in multi-agent reinforcement learning.

Our contributions: We identify two potential benefits of communication in these multi-agent systems. First, the instantaneous effect of communication is to enable better *coordination* among the agents in reaching their collaborative goal. This effect is in line with that commonly identified in the existing literature on multi-agent decision making.

Moreover, we identify a forward effect of communication, as it relates to agents' *learning*: we show that communication can also improve the learning of *both* agents. Such improvement may be expected in the less informed agent's learning, as the quality of information available to this agent improves through communication. More interestingly, our analysis shows that the learning of the informed agent also improves, even though communication does not affect the quality of the information available to this agent. That is, agents can benefit even if communication does not provide them with additional information. Intuitively, this finding can be explained as follows: by aiding the learning of another less informed agent, and given the collaborative nature of the agents' goal, an informed agent can collect more "informative" sample trajectories during its repeated interactions with the environment, hence enhancing its learning. Our analysis thus elaborates on the coupling between the agents' coordination and learning tasks in collaborative multi-agent environments.

We then show that the realization of the two identified benefits from communication will in general depend on its timing, and more specifically, on the agents' policy initialization. In particular, depending on the initialization of agents' policy parameters, communication in earlier stages of the game may in fact decrease the likelihood of agents' coordination and deter agents from learning. In these scenarios, and especially when agents are more shortsighted (i.e., place higher value on their immediate rewards), delayed information sharing may be preferred. We illustrate our findings through numerical examples.

Our main contributions can be summarized as follows:

- We show that the potential *benefits* of communication in multi-agent systems are in general two-fold: it not only facilitates *coordination*, but can further enhance the *learning* of *both* informed and less informed agents.
- We show that the realization of these benefits from communication is dependent on its *timing*: communication in earlier stages of the game may in fact *hinder* both coordination and agents' learning, making delayed communication preferable.
- We identify the *parameters affecting the optimal timing* of communication, including the policy initializations, the agents' patience, and the quality of the agents' independent observations.

The remainder of the paper is organized as follows. We present the model for the environment in Section II, followed by the multi-agent learning algorithm in Section III. Section IV analyzes the potential benefits of communication. Section V illustrates the effects of communication timing. We validate our results through numerical studies in Section VI, and conclude with a discussion of some implications of our findings in Section VII.

II. Model and Preliminaries

A. The POMDP environment

We consider a multi-agent Partially Observable Markov Decision Process (POMDP) in which N agents take actions over an infinite time horizon $t = \{1, 2, \ldots\}$. We begin by introducing the general model, followed by the specific parameters used in establishing our analytical results.

In general, a POMDP $(\mathcal{N}, \mathcal{A}, \mathcal{S}, p, \mathcal{O}, \mathbf{q}, \mathbf{r}, \delta)$ is determined by the following elements:

Agents: A set of agents $\mathcal{N}$ interact with one another, and with the environment.

Actions: At time t, each agent $i \in \mathcal{N}$ takes an action $a_{it} \in \mathcal{A}_i$, where $\mathcal{A}_i$ denotes the agent's discrete action space. Let $\mathbf{a}_t = \{a_{1t}, \ldots, a_{Nt}\}$ denote the vector of joint actions of all agents at time t, and $\mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_N$ denote the joint action space.

States: The underlying environment evolves according to a Markov Decision Process (MDP) with a finite state space $\mathcal{S} = \{s_1, \ldots, s_m\}$. The state of the MDP at time t is denoted $s_t \in \mathcal{S}$. Following agents' actions $\mathbf{a} \in \mathcal{A}$, the environment will transition from state s to s' according to a transition probability $p : \mathcal{S} \times \mathcal{A} \to \Pi(\mathcal{S})$. Denote $p(s', s, \mathbf{a}) := P(s_{t+1} = s' | s_t = s, \mathbf{a}_t = \mathbf{a})$.

Observations: In a POMDP, the state of the environment is not directly observed by the agents; rather, each agent i has a private observation or belief about s_t, denoted o_{it}. These private observations are generated according to an observation function $q_i : \mathcal{S} \to \mathcal{O}$, where $\mathcal{O}$ denotes the finite set of all possible observations. Denote $q_i(o, s) := P(o_{it} = o | s_t = s)$. We assume that the private observations $\{o_{it}, \forall i\}$ are independent across agents. Note also that the observation functions q_i are agent-dependent, so that the accuracy of the observations can vary across agents.

Rewards: Each agent i collects a reward r_{it} at each time t. The reward is determined by the reward function $r_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, which depends on the current state, as well as the choice of actions of all users. All agents discount future rewards by a factor δ. The discounted long run reward of agent i will be given by $R_i := \sum_t \delta^t r_{it}$. Each agent's goal is to maximize its own expected long-run reward $\mathbb{E}[R_i]$.

Based on the above definition, we observe that as the agents' rewards depend not only on the state of the environment and their own actions, but also on the actions of other agents, the POMDP can be viewed as an N-person game among the agents. The type of this game will be determined by the relation between the agents' reward functions. In particular, two special cases that are commonly of interest include *cooperative* ($r_i = r, \forall i$) and *competitive* or *zero-sum* ($r_1 = -r_2$) games. More generally, any environment in which $r_i \neq r_j$ for at least one pair of agents i, j is referred to as a *non-cooperative*

game. In this paper, our focus is on cooperative environments; the study of communication in non-cooperative games is an interesting direction of future work.

B. The collaborative binary POMDP

For concreteness, we will evaluate the effects of communication in the following POMDP. We will focus on an $N = 2$ agent cooperative environment with two states, two observations, and two actions.[1]

Specifically, we consider two agents, $i \in \mathcal{N} = \{1, 2\}$, selecting one of two possible actions $a_{it} \in \mathcal{A} = \{0, 1\}$ at each time step t. They interact in an environment with states $s_t \in \mathcal{S} = \{-1, 1\}$, and each obtain an observation $o_{it} \in \mathcal{O} = \{-1, 1\}$ of the state at each time step t.

To model heterogeneity in agents' observations, we let agent 1 be the more informed agent. In particular, we will proceed with the analysis under the following assumption. We let agent 1 be fully informed, i.e. $q_1(o_1 = s, s) = 1$ for all s. On the other hand, agent 2 is only partially informed, with $q_2(o_2 = s, s_2 = s) = \beta < 1$.[2] Finally, the fact that agent 1 is better informed is common knowledge between the agents.

Following agents' actions, we let the rewards be given by $r_i(s = 1, a_1 = 1, a_2 = 1) = 1$ and $r_i(s = -1, a_1 = 0, a_2 = 0) = 1$, for all i, with the remaining rewards being zero. Therefore, it is beneficial for the agents to coordinate (so as to take the same action) with the right coordination actions dependent on the current state (hence the need for learning). Finally, for the state transition probabilities, we assume the state will self-transition when agents correctly coordinate, but transition to the other state under other combinations of actions. That is, we let $p(s' = 1, s = 1, a_1 = 1, a_2 = 1) = 1$ and $p(s' = -1, s = -1, a_1 = 0, a_2 = 0) = 1$.

III. The Multi-Agent REINFORCE Algorithm

A. Multi-agent reinforcement learning

Each agent's goal when interacting with the POMDP environment is to choose her actions so as to maximize her expected long run reward $\mathbb{E}[R_i]$. The choice of actions is determined by the agent's policy. Specifically, an agent's (stochastic) policy $\pi_i : \mathcal{O} \to \Pi(\mathcal{A}_i)$ maps her private observation of the current state of the environment to the probability of selecting each action. If all of the environments' parameters where known to the agents, they could solve for the optimal policy using dynamic programming methods, and behave accordingly.

Nevertheless, when the environment is unknown to the agents, each agent only has access to her own rewards and private observations, collected through repeated interactions with the environment, while all else in unknown. These scenarios are the focus of the reinforcement learning (RL) literature. This literature studies how an agent should learn to act in such unknown environments, by using RL algorithms which repeatedly take the outcomes attained by following the agent's current policy as input, and output an updated policy accordingly. More generally, the multi-agent reinforcement learning literature (MARL) studies this problem in the setting where multiple agents simultaneously interact and learn their optimal polices in an unknown environment.

Here, following the literature on policy iteration in reinforcement learning, we assume that the agents choose the general form of their policies from a parameterized set $\{\pi(a|o, \theta)\}$, with the choice of parameter θ determining an agent's specific policy. In this approach, the policy updates can be done by adjusting only the parameter of the policy.

In particular, for the RL algorithm used by the agents, we will consider the well-known REINFORCE algorithm of [15]. REINFORCE, which is often considered a special case of actor-critic algorithms [16], [17], was originally proposed for single-agent reinforcement learning problems. Here, we present a variant with extension to multi-agent environments which incorporates communication.

B. The REINFORCE algorithm

We consider the episodic REINFORCE algorithm [15], also known as the Monte Carlo policy gradient algorithm [22], in which agents update the parameter of their policy based on their interactions with the environment over multiple episodes. Specifically, in an episode of length T, the agent uses her current policy to collect a set of observations, actions, and rewards $\{o_i, a_i, r_i\}_{i=1}^{T}$, and then runs the REINFORCE algorithm to update the parameter of her policy.

The update after the conclusion of each episode is as follows. Let $J_{\theta_i} = \mathbb{E}_{s\sim\pi_{\boldsymbol{\theta}}, \mathbf{a}\sim\pi_{\boldsymbol{\theta}}}[R_i]$ be the expected reward of agent i with respect to the state and action distribution realizations under a policy parametrized by $\boldsymbol{\theta}$. In the REINFORCE algorithm, agents use gradient ascent to update their policy parameter θ_i in the direction of the gradient of this reward. The policy gradient theorem [16] states that this gradient can be approximated by,

$$\nabla_\theta J_{\theta_i} \propto \mathbb{E}_{\pi_{\boldsymbol{\theta}}}[\nabla_\theta \log \pi(a|o, \theta) \sum_{l=t}^{T} \delta^{l-t} r_{il}] \ . \tag{1}$$

Therefore, using gradient ascent, the updated parameter at time $t + 1$ will be given by,

$$\theta_{i(t+1)} = \theta_{it} + \alpha \nabla_\theta J_{\theta_{it}} \ , \tag{2}$$

where α is the learning rate.[3] The steps of the algorithm are outlined in Algorithm 1.

For the parametrized family of policies $\{\pi_\theta\}$ to be used in our proposed POMDP, we will assume that agents are *Bernoulli-logistic units*.[4] This family is defined as follows:

[1] All restrictions of the size of the environment are without loss of generality to the obtained results, and are adopted to simplify the exposition. In particular, the extension to N agents is possible at the expense of additional notational complexity, and can be found in the online appendix [21].

[2] We will assume, without loss of generality, that $\beta \geq 0.5$.

[3] Throughout, we assume that the step size α is chosen appropriately to guarantee convergence.

[4] The choice between different policy parametrization options can be a way to inject prior knowledge in the learning algorithm. In particular, for the binary POMDP of Section II-B, the family of Bernoulli-logistic policies is particularly suitable as it closely approximates the optimal policy for large θ. We will further set the bias term of the units to zero for simplicity.

Algorithm 1 The Monte-Carlo policy gradient (REINFORCE) algorithm

function REINFORCE
Initialize θ_i arbitrarily
  for each episode $\{o_{it}, a_{it}, r_{it}\}_{t=1}^{T} \sim \pi(\cdot|\cdot, \theta_i)$ **do**
    for $t = 1 : T$ **do**
      $\theta_i \leftarrow \theta_i + \alpha \nabla_\theta \log \pi(a_{it}|o_{it}, \theta_i) \sum_{l=t}^{T} \delta^{l-t} r_{il}$
    end for
  end for
return θ_i
end function

given the current parameter θ_{it} in agent i's policy, she chooses her action a_{it} with the following probabilities:

$$\pi_{\theta_{it}}(a_{it}|o_{it}) = \begin{cases} \frac{1}{1+e^{-\theta_{it}o_{it}}} & \text{for } a_{it} = 1, \\ \frac{e^{-\theta_{it}o_{it}}}{1+e^{-\theta_{it}o_{it}}} & \text{for } a_{it} = 0. \end{cases} \quad (3)$$

For this class of policies, using the policy gradient theorem, the change in the policy parameter in the REINFORCE algorithm following step t is given by,

$$\Delta\theta_{it} = \alpha \mathbb{E}_\pi [\sum_{l=t}^{T} \delta^{l-t} r_{il} \cdot \begin{cases} \frac{o_{it}e^{-\theta_{it}o_{it}}}{1+e^{-\theta_{it}o_{it}}} & \text{for } a_{it} = 1 \\ \frac{-o_{it}}{1+e^{-\theta_{it}o_{it}}} & \text{for } a_{it} = 0 \end{cases}] . \quad (4)$$

To extend the above algorithm to multi-agent settings with communication, we assume that agents have access to a communication channel through which they can share their private observations with one another, and update the input to their policy according to the shared information.[5]

More specifically, for the environment of Section II-B, given our assumption that agent 1 is fully informed, she will have the ability to share any desired level of information with agent 2. Given that agent 2 is only partially informed, combined with the cooperative nature of the game, a natural conjecture is that *full information sharing and adoption* will lead to a Nash equilibrium of the cooperative game. Specifically, we may expect that agent 1 will fully share her state observation with agent 2, and agent 2 will discard his own observation and substitute agent 1's communicated observation as an input to his policy instead of his own observation.

In the next section, we show that the above can indeed be an equilibrium under an assumption on the initializations of the policies' parameters. We show how this equilibrium improves upon the outcomes from the default equilibrium in which agents learn independently. Through this analysis, we identify the benefits of communication in terms of both facilitating coordination and improving agents' learning.

IV. Benefits of Communication: Coordination and Learning

Consider the POMDP of Section II-B when agents learn using the REINFORCE algorithm of III-B. To find conditions under which full information sharing and adoption can be a Nash equilibrium in this setting, we need to evaluate the benefits of following the equilibrium strategies for both agents. We separate the analysis based on the instantaneous (current stage) and forward (long-run) effects of the shared information.

A. Instantaneous effect: improved coordination

The immediate effect of sharing information can be seen in the agents' expected instantaneous reward. Our first proposition formally analyzes this effect.

Proposition 1 (Information sharing aids coordination)**.** *Agents' expected (instantaneous) reward at time t is increasing in the information shared at time t if and only if agent 2's policy's parameter at time t is non-negative, i.e., $\theta_{2t} \geq 0$. In particular, when $\theta_{2t} \geq 0$, full information sharing by the informed agent, and full adoption by the less informed agent, will lead to the most increase in the instantaneous reward.*

The proof is given in the appendix. This result is intuitively interpreted as follows. Sharing of information from a more informed agent 1 to the less informed agent 2, and the adoption of this information by agent 2, will lead agent 2 to choose his action according to the correct state of the environment. This will lead to an increase in the expected reward from the current state *if and only if* the current policy of agent 2 is such that he is choosing the optimal action with higher frequency. In the POMDP of Section II-B, this is equivalent to having $\theta_{2t} \geq 0$.

In particular, one may envision scenarios in which the informed agent would be better off when delaying information sharing, so that the policy parameters have been improved over their random initialization, and therefore sharing of information can aid correct coordination. We elaborate on this effect further in Section V.

It is also worth noting that the statement of Proposition 1 is independent of the informed agent's policy parameter. This is expected as the sharing of information does not affect the informed agent's choice of action in the current step. Nonetheless, as we show in the next section, sharing of information will affect the choice of actions by the informed agent in *future* steps.

B. Forward effect: improved learning

The arguments presented above account only for the effects of the shared information on the agents' current reward, but not on the *future* behavior of the agents. In this section, we show that the shared information will affect the parameter updates of both agents, and consequently, all future rewards. More specifically, if communication between agents occurs at a time $1 \leq t_c \leq T$ during an episode, the collected traces $\{o_{it}, a_{it}, r_t\}_{t=t_c}^{T}$, and the REINFORCE updates at the end of the episode, will be affected by this communication. We therefore evaluate the effects of communication on both agents' parameter updates given these changes.

[5]Availability of a single communication channel entails public communication. Assuming pairwise communication channels which enable private communication may be of interest, but as our results show, will yield the same outcomes in cooperative environments.

Proposition 2 (Information sharing aids agents' learning)**.** *When the informed agent 1 shares her state observation with the less informed agent 2:*

- *If $\theta_{2t} \geq 0$ at the beginning of an episode, agent 1's learning improves.*
- *Agent 2's learning always improves.*

The proof is given in the appendix. It is worth noting that agent 2 always benefits from communication, while the same is not necessarily true for agent 1. To see why, note that by Proposition 1, the instantaneous reward collected at time t by agents increases if and only if $\theta_{2t} \geq 0$. This means that when $\theta_{2t} < 0$ at the beginning of a learning episode, communication decreases the likelihood that the agents collect non-zero rewards in that episode; however, non-zero rewards are the informative samples that guide the updates in the REINFORCE algorithm. Thus, communicating when $\theta_{2t} < 0$ decreases the likelihood that agent 1 collects informative traces, and hence is not necessarily beneficial to her learning. For agent 2 on the other hand, even though the likelihood of having informative traces decreases, communication allows this agent to associate the correct actions with the correct true state of the environment during the updates. Proposition 2 establishes that the latter effect of correct association is more important than collecting additional informative traces, and hence the less informed agent 2's learning always improves under communication.

Note also that the condition for improvement of agent 1's learning is only a sufficient condition. That is, it is still possible for agent 1 to benefit from communication even if $\theta_{2t} < 0$ at the beginning of the REINFORCE episode. This is because by improving agent 2's learning (even if collecting less informative traces in the current episode), agent 1 can increase agent 2's learning speed, and hence, the likelihood that they can collect higher rewards and more informative samples in future episodes. We elaborate on the tradeoffs between these effects in the next section.

V. Communication Timing: the Effects of Policy Initialization

As shown in Propositions 1 and 2, if the parameter policy θ_2 is initialized to a positive value, information sharing is always beneficial to both agents. Nevertheless, with a negative initialization of the less informed agent 2's policy, $\theta_2 < 0$, sharing of information about the state of the environment will in fact reduce the agents' expected instantaneous reward, and degrade agent 1's learning, as long as θ_{2t} remains negative.

On the other hand, communication always improves agent 2's learning, which can in turn improve the future rewards and learning of both agents given the collaborative nature of their goal. Note also that through the use of the REINFORCE algorithm, θ_{2t} will improve over its initial (negative) initialization; once θ_{2t} becomes non-negative, information sharing will become beneficial to both agents. Given this, it will ultimately benefit agent 1 to initiate communication at some point in the interaction, once learning has progressed enough.

That being said, the optimal range of policy parameters is in general not known a priori in learning problems, and can therefore not be used directly to determine the optimal timing of communication. In this section, we identify other parameters that affect the (sub-)optimality of early communication, and can therefore be used to guide the decision of when to communicate.

A. Parameters affecting the optimal timing of communication

We illustrate the trade-offs between the instantaneous and forward effects of communication in a minimal instance of the binary collaborative POMDP consisting of a 2 episodes of length $T = 1$, in which policy parameters are initialized arbitrarily. We compare agent 1's expected reward when starting communication at the beginning of the game, or when delaying communication until the beginning of the second episode.

Denote the instantaneous reward at time t by r_t^o, and the policy parameters of agent i at time t by θ_{it}^o, where $o \in \{C, D\}$ denotes the decision to communicate or delay, respectively. Early communication will be beneficial if and only if,

$$\Delta\mathbb{E}(r_1) + \delta\Delta\mathbb{E}(r_2) \geq 0 \ . \tag{5}$$

where $\Delta\mathbb{E}(r_1) := \mathbb{E}[r_1^C - r_1^D]$ is the instantaneous effect, and $\Delta\mathbb{E}(r_2) := \mathbb{E}[r_2^C - r_2^D]$ is the forward effect of communication.

a) Instantaneous effect: From the proof of Proposition 1, we know that agents' reward at t is,

$$E[r_t] = \frac{e^{\theta_{1t}}}{1+e^{\theta_{1t}}}\left(\frac{e^{\theta_{2t}}}{1+e^{\theta_{2t}}}\beta + \frac{e^{-\theta_{2t}}}{1+e^{-\theta_{2t}}}(1-\beta)\right) \ . \tag{6}$$

Then,

$$\Delta\mathbb{E}(r_1) = (1-\beta)\frac{e^{\theta_{11}}}{1+e^{\theta_{11}}}\frac{e^{\theta_{21}} - e^{-\theta_{21}}}{(1+e^{\theta_{21}})(1+e^{-\theta_{21}})} \ . \tag{7}$$

b) Forward effect: We next look at the parameter updates to be used in the second episode, θ_{12} and θ_{22}, with and without communication. From the proof of Proposition 2, we have,

$$\begin{aligned}\theta_{12} - \theta_{11} &= \alpha\frac{e^{\theta_{11}}}{(1+e^{\theta_{11}})^2}\frac{\beta e^{\theta_{21}} + (1-\beta)e^{-\theta_{21}} + 1}{(1+e^{\theta_{21}})(1+e^{-\theta_{21}})} \ , \\ \theta_{22} - \theta_{21} &= \alpha\frac{e^{\theta_{11}}}{1+e^{\theta_{11}}}\frac{2\beta - 1}{(1+e^{\theta_{21}})(1+e^{-\theta_{21}})} \ .\end{aligned} \tag{8}$$

Substituting for $\beta = 1$ in the above expressions determines the parameter updates θ_{i2}^C attained if communication happens in the first episode.

We are now ready to find $\Delta\mathbb{E}(r_2)$. As agent 1 is assumed to start sharing information at the second episode (i.e., $\beta = 1$), using (6), the change in her expected reward at time 2 will be,

$$\Delta\mathbb{E}[r_2] = \frac{1}{1+e^{-\theta_{12}^C}}\frac{1}{1+e^{-\theta_{22}^C}} - \frac{1}{1+e^{-\theta_{12}^D}}\frac{1}{1+e^{-\theta_{22}^D}} \ . \tag{9}$$

Substituting for (7) and (9) in (5), we conclude that communication in the first step is preferred over delayed communication if and only if,

$$\begin{aligned}&(1-\beta)\frac{e^{\theta_{21}} - e^{-\theta_{21}}}{(1+e^{-\theta_{11}})(1+e^{\theta_{21}})(1+e^{-\theta_{21}})} + \\ &\delta\left(\frac{1}{1+e^{-\theta_{12}^C}}\frac{1}{1+e^{-\theta_{22}^C}} - \frac{1}{1+e^{-\theta_{12}^D}}\frac{1}{1+e^{-\theta_{22}^D}}\right) \geq 0 \ .\end{aligned} \tag{10}$$

From the above, we make the following observation:

Proposition 3 (Parameters affecting the optimal timing of communication)**.** *The preferred timing of communication by agent 1 depends on the initialization of agent 2's policy parameter θ_2, the agents' patience δ (discounting of future rewards), and the quality of agent 2's observations in absence of communication β. In particular, for the problem instance of this section,*

1) ***[Initialization.]*** *If $\theta_{21} \geq 0$, it is optimal for agent 1 to start communication at the first episode.*
2) ***[Discounting of future rewards.]*** *If $\theta_{21} < 0$, there exists a $\delta_0 \in [0, 1)$, such that for $0 \leq \delta \leq \delta_0$, delayed sharing is preferred by agent 1.*
3) ***[Observation quality.]*** *When $\theta_{21} < 0$, there exists a $\beta_0 \in [\frac{1}{2}, 1)$, such that for $\frac{1}{2} \leq \beta \leq \beta_0$, delayed sharing is preferred by agent 1.*

The first statement above is consistent with Propositions 1 and 2, and can be seen by noting that both the (instantaneous and forward) terms in (10) are positive at $\theta_{21} \geq 0$.

For the second statement, we note that the first term of (10) (instantaneous effects of communication) is negative. The second term may be negative as well, as by (8), we will have $\theta_{12}^C \leq \theta_{12}^D$, that is, agent 1's learning degrades under communication. The second term (forward effect) overall can still be positive, as agent 2's learning can improve. Nonetheless, it is possible to find a small enough δ, such that (10) is dominated by the first term, and delayed communication is preferable.

Finally, for the third statement, we note the two-fold effect of agent 2's observation quality. First, for small β and negative initialization of θ_{21}, by (6), the loss of revenue due to miscoordination will be larger. On the other hand, increasing β from a small value up to 1 will lead to a more considerable improvement in θ_{21}. For the instance considered in this section, if agent 2 is sufficiently uninformed (smaller β), the loss of revenue dominates the benefit from improved learning. That is, perhaps surprisingly, delayed information has become more preferable even though the less informed agent's observations are of particularly low quality.

B. When to communicate?

From the analysis in the previous section, we have observed that in general, the policy initializations, the quality of observations in the absence of communication, and the agents' patience, all affect the optimal timing of communication.

First, note that the differentiation of the policy initialization in Proposition 3, which is based on the *sign* of the parameters, is indeed specific to the problem of Section II-B. More generally, our insights point to the fact that communication will be beneficial throughout the agents' interaction when the policy parameters are such that the optimal action is already being selected with a higher frequency. If this condition holds, sharing of information will not misguide the action choice of the less informed agent, and can hence aid the coordination and learning of the agents.

The optimal range of policy parameters is nonetheless not known a priori in learning problems, and can therefore not be used directly to determine the optimal timing of communication. A possible proxy for evaluating the progress of learning is to keep track of the rate of change in the policy parameters: if the learning step size is chosen appropriately, policy gradient methods will lead to smaller updates as learning progresses. Communication can therefore begin once agents are sufficiently confident about their policies based on the rate of change in their policy parameters.

Finally, for sufficiently patient agents ($\delta \to 1$), communication from early stages of the game will always be beneficial. This is because through the use of learning algorithms, even with suboptimal initializations, the policy parameters will gradually improve towards their optimal values (and the improvement will be faster under communication). Therefore the agents will reap the benefits of communication sooner through adapting early information sharing, at the expense of lower rewards from a limited number of earlier stages.

VI. Numerical Examples

A. Benefits of communication

We begin by illustrating the benefits of communication by comparing the outcomes of the agents' interactions, specifically their expected discounted rewards and the progress of their learning, with and without communication. For this part, we will initialize both θ_1 and θ_2 randomly to a non-negative value between $[0, 1]$. In addition, we let $\delta = 0.9$, $\beta = 0.7$, and $\alpha = 0.1$. Figures 1 and 2 illustrate the learned parameters and expected rewards after 20 episodes of length $T = 10$, with and without communication, respectively. The results are averaged over random initializations of the starting state and parameter policies over $10,000$ trials.

We observe that as illustrated in Fig. 1, communication indeed aids the learning of both agents (Proposition 2). First, it is worth noting that, without communication, the more informed agent 1 learns faster than the less informed agent 2. This is due to the fact that, without information sharing, agent 2 at times associates his updates to an incorrect state due to the imperfect observations of the environment. On the other hand, the policies learned in the presence of communication outperform those of *both* agents in the absence of communication. That is, information sharing aids the learning of both the better informed and the less informed agent. Note also that under full sharing and adoption, both agents will perform the same updates, and as their starting parameter is equal on average, their policy parameter curves overlap throughout. Lastly, as shown in Fig. 2, both agents will indeed benefit from communication due to the increase in their expected rewards. Note that by the cooperative nature of the game, both agents receive the same rewards.

B. Timing of communication

Next, we illustrate the findings of Section V. For this part, we will again initialize θ_1 randomly to a value between $[0, 1]$, but will set θ_2 to a negative value in $[-1, 0]$. We let

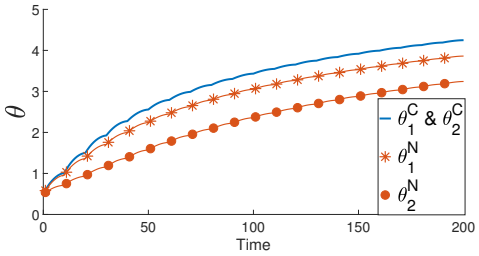

Fig. 1. Information sharing improves the speed of learning of both agents (under positive parameter initializations).

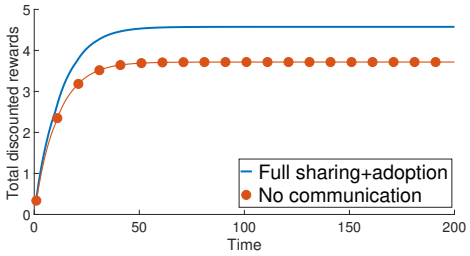

Fig. 2. Information sharing improves the expected rewards of both agents (under positive parameter initializations).

$\delta = 0.2$ (i.e., less patient agents), $\beta = 0.7$, and $\alpha = 0.1$. Figures 3 and 4 illustrate the learned parameters and expected rewards after 20 episodes of length $T = 10$, with and without communication, respectively. The results are averaged over random initializations of the starting state and parameter policies over $10,000$ trials.

We first note that as shown in Fig. 3, communication will, in the long-run, improve the learning of both agents. Nevertheless, in the early steps, agent 1's learning parameter without communication θ_1^N, surpasses that in the presence of communication θ_1^C. This is due to the negative initialization of agent 2's policy parameter, which causes the shared information to misguide his actions, hence reducing coordination and hindering agent 1's learning. As illustrated in Fig. 4, the expected rewards of agents is in fact higher *without* communication, due to the miscoordination caused by communication in the early stages. Note also that, as shown in Proposition 2, information sharing is beneficial to the less informed agent 2's learning irrespective of the policy initialization.

VII. Discussion and Conclusion

We studied the problem of concurrent learning and coordination of two heterogeneous agents in a partially observable environment, when a better informed agent can share her information with a less informed agent. We formally analyzed the benefits of incorporating this explicit communication in agents' learning algorithm, and showed that information sharing can enhance coordination and also improve the learning of both agents in the long-run, but that it may hinder coordination and learning in early stages of the cooperative game depending on the initialization of agents' policies.

A main implication of our findings is in the design and operation of self-organizing multi-agent systems in unknown environments, and in particular those based on the edge/fog computing paradigm. In this framework, learning and control are performed primarily by agents residing on the edges of the network. As a result, agents refine their policies and consequently select their actions in a decentralized manner. Our results show that leveraging cloud connectivity for limited communication can be beneficial for multi-agent learning and coordination, given correct timing or when agents are sufficiently patient.

It is worth mentioning that our choice of allowing agents to communicate only local observations (rather than, e.g., policies/actions) is motivated not only by the latency and costs of communication, but also by the possibility that agents may in general use heterogeneous learning algorithms and policies, so that communication of information on policies may not be interpretable by all agents. This type of communication is particularly relevant in edge/fog computing scenarios, where heterogeneous policies may be employed by different devices as a result of their computational constraints.

Main directions of future work include analyzing the effects of communication in alternative reinforcement learning algorithms, including analytical evaluation of the benefits and drawbacks of communication in other actor-critic algorithms, as well as empirical evaluation of the effects of communication timing on the performance of multi-agent deep reinforcement learning algorithms.

APPENDIX

A. Proof of Proposition 1

The expected reward of the agents is the same at time t, and is given by,

$$\mathbb{E}_{s,\pi}[r(s,a_{1t},a_{2t})] = P(s_t=-1)E[r(-1,a_{1t},a_{2t})|s_t=-1] + P(s_t=+1)E[r(1,a_{1t},a_{2t})|s_t=+1]. \quad (11)$$

We analyze case $s_t = -1$; a similar argument holds for $s_t = 1$.

Recall that in state $s_t = -1$, agents will obtain a non-zero reward if and only if $a_{1t} = a_{2t} = 0$. Therefore,

$$\begin{aligned} E[r(-1,a_{1t},a_{2t})|s_t=-1] &= \pi_{\theta_{1t}}(a_{1t}=0|o_{1t}=-1) \\ &\quad (\pi_{\theta_{2t}}(a_{2t}=0|o_{2t}=-1)q_2(o_{2t}=-1|s_t=-1) \\ &\quad + \pi_{\theta_{2t}}(a_{2t}=0|o_{2t}=1)q_2(o_{2t}=1|s_t=-1)) \\ &= \frac{e^{\theta_{1t}}}{1+e^{\theta_{1t}}}\left(\frac{e^{\theta_{2t}}}{1+e^{\theta_{2t}}}\beta + \frac{e^{-\theta_{2t}}}{1+e^{-\theta_{2t}}}(1-\beta)\right), \end{aligned} \quad (12)$$

where we have used the fact that agent 1 knows the state accurately (i.e., $o_{1t} = -1$ w.p. 1), and agent 2's knowledge is given by $\beta = q_2(o_{2t} = -1, s_t = -1)$. Define,

$$h(\beta,\theta_{2t}) := \beta\left(\frac{e^{\theta_{2t}}}{1+e^{\theta_{2t}}} - \frac{e^{-\theta_{2t}}}{1+e^{-\theta_{2t}}}\right) + \frac{e^{-\theta_{2t}}}{1+e^{-\theta_{2t}}}. \quad (13)$$

Note that $h(\cdot)$ is a non-decreasing function of β if and only if $\frac{e^{\theta_{2t}}-e^{-\theta_{2t}}}{(1+e^{\theta_{2t}})(1+e^{-\theta_{2t}})} \geq 0$, which happens if and only if $\theta_{2t} \geq 0$.

That is, an increase in β will increase agents' reward if and only if $\theta_{2t} \geq 0$. In particular, full information sharing by the

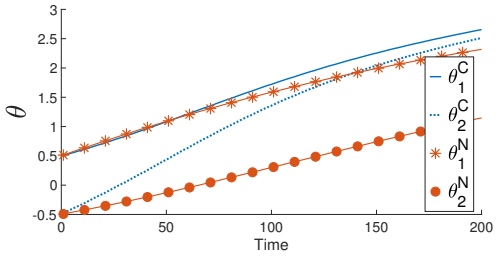

Fig. 3. Under negative parameter initialization for the less informed agent, information sharing will ultimately improve the learning of both agents, but may cause a slow down in the learning of the informed agent at early stages.

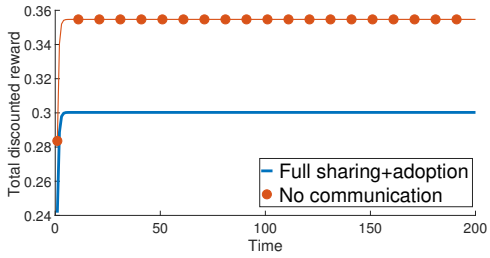

Fig. 4. Under negative parameter initialization for the less informed agent, information sharing is not necessarily beneficial due to reduced coordination and slow down of learning in early stages.

informed agent, which under full adoption by the less informed agent leads to $\beta = 1$, would lead to the most increase in the agents' instantaneous reward. ■

B. Proof of Proposition 2

We first note that for the environment of Section II-B, agents' optimal policy parameter is $\theta \to +\infty$. As a result, we establish improved learning by showing that agents take larger gradient steps under communication.

We consider an episode of length T, and assume that communication occurs (only) at some time $1 \le t_c \le T$ within the episode. We establish the effects of this change on the parameter updates of both agents. The same analysis can be carried out if information sharing occurs at multiple steps in the episode, as the effects are superimposed.

Recall that under the REINFORCE algorithm, the change in agent i's parameter based on step t's action and reward is,

$$\Delta\theta_{it} = \alpha\mathbb{E}_\pi[\sum_{l=t}^{T}\delta^{l-t}r_{il}\cdot\begin{cases}\frac{o_{it}e^{-\theta_{it}o_{it}}}{1+e^{-\theta_{it}o_{it}}} & \text{for } a_{it}=1\\ \frac{-o_{it}}{1+e^{-\theta_{it}o_{it}}} & \text{for } a_{it}=0\end{cases}] \ . \quad (14)$$

We will evaluate agents' parameter updates at a given step $t_c \le t \le T$. Note that the final update at time T will lead to the parameter $\theta_{i(T+1)}$, which is the policy initialization at the beginning of the next episode, and hence, determines the new policy based on which agent i collects rewards in the future.

We start with agent 2, and separate the expression based on the realization of the state s_t. The change in agent 2's policy parameter when $s_t = -1$ is given by,

$$\theta_{2(t+1)} - \theta_{2t} = \alpha\cdot$$
$$\mathbb{E}[(r(s_t=-1,a_{1t},a_{2t}) + \sum_{l=t+1}^{T}\delta^{l-t}r(s_l,a_{1l},a_{2l}))\cdot$$
$$\begin{cases}\frac{o_{2t}e^{-\theta_{2t}o_{2t}}}{1+e^{-\theta_{2t}o_{2t}}} & \text{for } a_{2t}=1\\ \frac{-o_{2T}}{1+e^{-\theta_{2t}o_{2t}}} & \text{for } a_{2t}=0\end{cases}|s_t=-1] \ . \quad (15)$$

From the above, we note that as the agents take their sum reward looking forward in determining each REINFORCE update, the update at time t will depend on the realization of actions and rewards in the trace collected in the future up to time T. At the same time, the realization of the state at time $t+1$ (and hence, forward) will itself depend on the actions taken in the current state. We therefore have four different possible updates, depending on the profile of actions at time t and the realization of state s_{t+1}. Let $\bar{R}(s_{t+1}=s) := \mathbb{E}[\sum_{l=t+1}^{T}\delta^{l-t}r(s_l,a_{1l},a_{2l})|\ s_{t+1}=s]$.

For the action profile $\mathbf{a}_t = (0,0)$, the state will evolve to $s_{t+1} = -1$, and we have the following update:

$$\Delta(0,0) := \alpha\cdot(1+\bar{R}(s_{t+1}=-1))(\pi_{\theta_{1t}}(a_{1t}=0|o_{1t}=-1)$$
$$(\pi_{\theta_{2t}}(a_{2t}=0|o_{2t}=-1)q_2(o_{2t}=-1|s_t=-1)\frac{1}{1+e^{\theta_{2t}}}+$$
$$\pi_{\theta_{2t}}(a_{2t}=0|o_{2t}=1)q_2(o_{2t}=1|s_t=-1)\frac{-1}{1+e^{-\theta_{2t}}}))$$
$$= \alpha(1+\bar{R}(s_{t+1}=-1))\frac{e^{\theta_{1t}}}{1+e^{\theta_{1t}}}f(\beta,\theta_{2t}) \ , \quad (16)$$

where,

$$f(\beta,\theta_{2t}) := \frac{2\beta-1}{(1+e^{\theta_{2t}})(1+e^{-\theta_{2t}})} \ . \quad (17)$$

We observe that $f(\beta,\theta_{2t})$ is increasing in β.

For all other action profiles, the agents will not receive any reward from step t, and the state will transition to $s_{t+1} = +1$. Following steps similar to (16), the updates for these action profiles are given by:

$$\Delta(1,0) := \alpha\bar{R}(s_{t+1}=+1)\frac{1}{1+e^{\theta_{1t}}}f(\beta,\theta_{2t}) \ ,$$
$$\Delta(0,1) := -\alpha\bar{R}(s_{t+1}=+1)\frac{e^{\theta_{1t}}}{1+e^{\theta_{1t}}}f(\beta,\theta_{2t}) \ , \quad (18)$$
$$\Delta(1,1) := -\alpha\bar{R}(s_{t+1}=+1)\frac{1}{1+e^{\theta_{1t}}}f(\beta,\theta_{2t}) \ .$$

Putting these expressions together, leads to,

$$\theta_{2(t+1)} - \theta_{2t} = \alpha(1+\bar{R}(s_{t+1}=-1) - \bar{R}(s_{t+1}=+1))$$
$$\frac{e^{\theta_{1t}}}{1+e^{\theta_{1t}}}f(\beta,\theta_{2t}) \ . \quad (19)$$

First note that $f(\beta,\theta_{2t})$ is non-decreasing in β. We also note that the terms $\bar{R}(s_{t+1}=-1)$ and $\bar{R}(s_{t+1}=+1)$ are evaluated based on the trace collected from time $t+1$ onwards, and are independent from communication at time t. Further, as they are generated using the same, they are in fact equal in expectation. We conclude that agent 2's update in (19) is increasing in β, and is maximized at $\beta = 1$, i.e., when full information

is shared by agent 1 and adopted by agent 2. The argument for starting from $s_t = +1$ is similar. Therefore, information sharing benefits agent 2's learning.

We now turn to agent 1, and consider the update at time t, separating the expression based on the state at time t. The change in agent 1's policy parameter when $s_t = -1$ is,

$$\theta_{1(t+1)} - \theta_{1t} = \alpha \cdot \mathbb{E}[(r(s_t = -1, a_{1t}, a_{2t}) + \sum_{l=t+1}^{T} \delta^{l-t} r(s_l, a_{1l}, a_{2l})) \cdot \begin{cases} \frac{o_{1t} e^{-\theta_{1t} o_{1t}}}{1+e^{-\theta_{1t} o_{1t}}} & \text{for } a_{1t} = 1 \\ \frac{-o_{1t}}{1+e^{-\theta_{1t} o_{1t}}} & \text{for } a_{1t} = 0 \end{cases} | s_t = -1] \ . \quad (20)$$

From the above, we again note there are four different possible updates, depending on the profile of actions at time t and the realization of state s_{t+1}.

For the action profile $\mathbf{a}_t = (0, 0)$, the state will evolve to $s_{t+1} = -1$, the agents will collect a reward of 1, and we have the following update:

$$\begin{aligned} &\Delta(0,0) := \alpha(1 + \bar{R}(s_{t+1} = -1)) \\ &(\pi_{\theta_{1t}}(a_{1t} = 0 | o_{1t} = -1) \cdot \frac{1}{1 + e^{\theta_{1t}}} \cdot \\ &\quad (\pi_{\theta_{2t}}(a_{2t} = 0 | o_{2t} = -1) q_2(o_{2t} = -1 | s_t = -1) + \\ &\quad \pi_{\theta_{2t}}(a_{2t} = 0 | o_{2t} = +1) q_2(o_{2t} = 1 | s_t = -1)) \\ &= \alpha(1 + \bar{R}(s_{t+1} = -1)) \frac{e^{\theta_{1t}}}{(1 + e^{\theta_{1t}})^2} h(\beta, \theta_{2t}) \ , \end{aligned} \quad (21)$$

where $h(\beta, \theta_{2t})$ is defined in (13).

For all other action profiles, the agents will not receive any reward from step t, and the state will transition to $s_{t+1} = +1$. Following steps similar to (21), the updates for these action profiles are given by:

$$\begin{aligned} \Delta(1,0) &:= -\alpha \bar{R}(s_{t+1} = +1) \frac{e^{\theta_{1t}}}{(1 + e^{\theta_{1t}})^2} h(\beta, \theta_{2t}) \ , \\ \Delta(0,1) &:= \alpha \bar{R}(s_{t+1} = +1) \frac{e^{\theta_{1t}}}{(1 + e^{\theta_{1t}})^2} (1 - h(\beta, \theta_{2t})) \ , \\ \Delta(1,1) &:= -\alpha \bar{R}(s_{t+1} = +1) \frac{e^{\theta_{1t}}}{(1 + e^{\theta_{1t}})^2} (1 - h(\beta, \theta_{2t})) \ . \end{aligned} \quad (22)$$

Putting these expressions together, leads to,

$$\theta_{1(t+1)} - \theta_{1t} = \alpha(1 + \bar{R}(s_{t+1} = -1) - \bar{R}(s_{t+1} = +1)) \frac{e^{\theta_{1t}}}{(1 + e^{\theta_{1t}})^2} h(\beta, \theta_{2t}) \ . \quad (23)$$

We first note that the first term above is again non-negative. In addition, we know that $h(\beta, \theta_{2t})$ is non-decreasing in β if and only if $\theta_{2t} \geq 0$. The argument when starting from $s_t = +1$ follows similar steps. We therefore conclude that sharing of information helps agent 1 take improved (here larger) gradient steps during the execution of the REINFORCE algorithm only in steps which $\theta_{2t} \geq 0$.

Lastly, note that h is increasing in θ_{2t}. Combined with the analysis of agent 2's learning, this means that while communication will not improve agent 1's learning at time t when $\theta_{2t} < 0$, it will still lead to a faster increase in $\theta_{2(t+1)}$, and could therefore improve agent 1's future updates. ■

REFERENCES

[1] H. Liu, B. Krishnamachari, and Q. Zhao, "Cooperation and learning in multiuser opportunistic spectrum access," in *Proc. IEEE ICC'08 Workshops*, May 2008.

[2] A. Anandkumar, N. Michael, and A. Tang, "Opportunistic spectrum access with multiple users: Learning under competition," in *Proc. IEEE INFOCOM'10*, March.

[3] Z. Han, D. Niyato, W. Saad, T. Başar, and A. Hjørungnes, *Game theory in wireless and communication networks: theory, models, and applications*. Cambridge University Press, 2012.

[4] W. Saad, Z. Han, H. V. Poor, and T. Basar, "Game-theoretic methods for the smart grid: An overview of microgrid systems, demand-side management, and smart grid communications," *IEEE Signal Processing Magazine*, vol. 29, no. 5, pp. 86–105, 2012.

[5] W. Ren and N. Sorensen, "Distributed coordination architecture for multi-robot formation control," *Robotics and Autonomous Systems*, vol. 56, no. 4, pp. 324–333, 2008.

[6] Y. Cao, W. Yu, W. Ren, and G. Chen, "An overview of recent progress in the study of distributed multi-agent coordination," *IEEE Transactions on Industrial informatics*, vol. 9, no. 1, pp. 427–438, 2013.

[7] X. He, H. Dai, and P. Ning, "Improving learning and adaptation in security games by exploiting information asymmetry," in *Proc. IEEE INFOCOM'15*, April 2015.

[8] "Fleet of sailboat drones could monitor climate changes effect on oceans," http://www.sciencemag.org/news/2018/03/fleet-sailboat-drones-could-monitor-climate-change-s-effect-oceans.

[9] "High above, drones keep watchful eyes on wildlife in africa," https://www.nytimes.com/2017/03/13/science/drones-africa-poachers-wildlife.html.

[10] "Honda unveils prototype e2-dr disaster response robot," https://spectrum.ieee.org/automaton/robotics/humanoids/iros-2017-honda-unveils-prototype-e2dr-disaster-response-robot.

[11] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Trans. Systems, Man, and Cybernetics, Part C*, vol. 38, no. 2, pp. 156–172, 2008.

[12] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," *IEEE Internet of Things J.*, vol. 3, no. 6, pp. 854–864, December 2016.

[13] Y. Xiao and M. Krunz, "QoE and power efficiency tradeoff for fog computing networks with fog node cooperation," in *Proc. IEEE INFOCOM'17*, May 2017.

[14] L. Tong, Y. Li, and W. Gao, "A hierarchical edge cloud architecture for mobile computing," in *Proc. IEEE INFOCOM'16*, May 2016.

[15] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," in *Reinforcement Learning*. Springer, 1992, pp. 5–32.

[16] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in neural information processing systems*, 2000.

[17] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," in *Advances in neural information processing systems*, 2000, pp. 1008–1014.

[18] J. K. Gupta, M. Egorov, and M. Kochenderfer, "Cooperative multi-agent control using deep reinforcement learning," in *International Conference on Autonomous Agents and Multiagent Systems*. Springer, 2017.

[19] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," *arXiv preprint arXiv:1705.08926*, 2017.

[20] R. Lowe, Y. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Advances in Neural Information Processing Systems*, 2017.

[21] "Online Appendix: Hurts to Be Too Early: Benefits and Drawbacks of Communication in Multi-Agent Learning," https://goo.gl/LrfMDW.

[22] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. The MIT Press, 2018. [Online]. Available: http://incompleteideas.net/book/the-book-2nd.html