Intelligent Communication over Realistic Wireless Networks in Multi-Agent Cooperative Games

Extended Abstract

Diyi Hu
University of Southern California
Los Angeles, CA, Unites States
diyihu@usc.edu

Chi Zhang
University of Southern California
Los Angeles, CA, Unites States
zhan527@usc.edu

Viktor Prasanna
University of Southern California
Los Angeles, CA, Unites States
prasanna@usc.edu

Bhaskar Krishnamachari
University of Southern California
Los Angeles, CA, Unites States
bkrishna@usc.edu

ABSTRACT

In MARL, communication among agents is essential to establish cooperation. Over the realistic wireless network, many factors can affect transmission reliability, especially considering that the wireless network condition varies with agents' mobility. We propose a framework that improves the intelligence of communication over *realistic* wireless networks in two fundamental aspects: (1) *When*: Agents learn the timing of communication based on message importance and wireless channel condition. We further propose a communication lagging technique to make the training end-to-end differentiable. (2) *What*: Agents augment message contents with wireless network measurements. The messages improve both the game and communication actions of the agents. Experiments on a standard environment show that compared with state-of-the-art, our framework enables more intelligent collaboration and thus achieves significantly better game performance, convergence speed and communication efficiency.

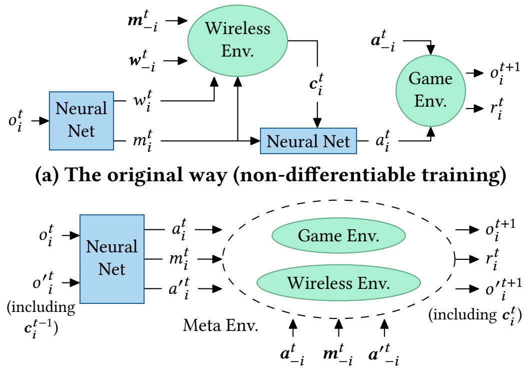
KEYWORDS

Multi-agent Reinforcement Learning; Wireless Communication

ACM Reference Format:
Diyi Hu, Chi Zhang, Viktor Prasanna, and Bhaskar Krishnamachari. 2022. Intelligent Communication over Realistic Wireless Networks in Multi-Agent Cooperative Games: Extended Abstract. In *Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), Online, May 9–13, 2022*, IFAAMAS, 3 pages.

1 INTRODUCTION

In Multi-Agent Reinforcement Learning (MARL), communication is critical to facilitate knowledge sharing and collaboration [3, 5, 17]. Most of the existing works [2, 11, 12, 15, 16, 18, 20] are based on unrealistic modeling of the wireless network environment and assume perfect transmission. However, in many practical MARL applications (especially those involving navigation: *e.g.*, fire fighting [7], Search-And-Rescue (SAR) [13]), agents communicate via a mobile ad-hoc network (MANET) [1, 9]. In such *realistic* wireless networks, successful transmission is not guaranteed due to limited bandwidth, signal path loss and fading, medium contention, interference, etc.. Moreover, the mutual influence between the game and wireless environments results in the *environment coupling* challenge, hampering the learning of communication strategies under realistic wireless networks: On the one hand, agents' mobility in the game environment leads to dynamic network connectivity and agents' communication actions have significant impact on medium contention and signal interference. On the other hand, changes in the wireless environment can affect agents' game actions (e.g., approaching others for better signal strength). We propose a general and flexible framework to enable intelligent multi-agent communication in realistic wireless networks from two fundamental aspects: **When**: We propose a "1-step communication lagging" trick to enable the formulation of "meta-environment". We further augment the action space so that communication scheme can be learned end-to-end via back-propagation. **What**: We augment the observation space by including wireless network measurements. Agents exchange such observations with small communication overhead. The proposed framework significantly improves existing MARL algorithms (both on- and off-policy ones) in a plug-and-play fashion.



(a) The original way (non-differentiable training)

(b) The new way (after lagging and alignment)

Figure 1: Two ways of agent-environment interaction

2 METHOD

Use subscript "$-i$" to denote variables from all agents other than i. A natural design, followed by [5, 15, 16], is to send and receive a message within the same step. In Figure 1a, at the beginning of step t, agent i makes local observation o^t_i on the game environment.

Proc. of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2022), P. Faliszewski, V. Mascardi, C. Pelachaud, M.E. Taylor (eds.), May 9–13, 2022, Online.

From o_i^t, a neural network generates a message m_i^t with weight w_i^t measuring message importance / relevance. Then based on w_i^t and $\boldsymbol{w}_{-i}^t$, the messages m_i^t and $\boldsymbol{m}_{-i}^t$ contend to go through the wireless channel. Since some transmissions may fail, we use $\boldsymbol{c}_i^t$ to denote the messages successfully received by i. Next, agent i's own message and all its received ones go through another neural network to generate the next action a_i^t. Finally, all agents interact with the game environment via a_i^t and $\boldsymbol{a}_{-i}^t$. The game environment returns the current reward r_i^t and next observation o_i^{t+1}. This concludes step t. A major drawback of the Figure 1a setting is training non-differentiability. When optimizing policy π_θ via back-propagation, the gradients of the parameters θ need to flow from a_i^t back to o_i^t. However, the wireless environment lies in the middle of step t. The mapping implemented by the realistic wireless channel, $\boldsymbol{m}^t, \boldsymbol{w}^t \rightarrow \boldsymbol{c}_i^t$, is non-differentiable [11]. We propose *"communication lagging"* to make training differentiable. Denote a' as communication actions and o' as wireless observations. In Figure 1b, an agent observes at the beginning of step t. Yet it does not generate m_i^t until the end of step t. After lagging, the "message-wireless environment" and "agent-game environment" interactions happen simultaneously (in dotted circle). Thus, the *alignment* of the two environments makes training fully differentiable. More importantly, it leads to a "meta-environment" abstraction, where the Markov Decision Process (MDP) controlling agents' behaviors can be reformulated, and both the action and observation spaces can be expanded.

We view the *meta-environment* as the *new* game environment: any algorithm on the original game environment can be directly applied on the meta-environment. To define agents' interaction with the meta-environment, we reformulate the MDP with expanded state and action spaces: $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \Omega, \mathcal{O}, \mathcal{R}, \gamma)$. Denote $\mathcal{S}$ as the state of the meta-environment. The *augmented* action space $\mathcal{A}$ consists of the game actions $\mathcal{A}^{\mathcal{T}}$ and communication actions $\mathcal{A}^{C}$. *i.e.*, $\mathcal{A} = \mathcal{A}^{\mathcal{T}} \times \mathcal{A}^{C}$. The *augmented* observation space Ω consists of the game observations $\Omega^{\mathcal{T}}$ and wireless observations Ω^{C}. *i.e.*, $\Omega = \Omega^{\mathcal{T}} \times \Omega^{C}$. $\mathcal{T}$ and $\mathcal{O}$ are defined on the meta-state and augmented observation and action spaces. Under the reformulated MDP, agents can learn a policy on communication actions to decide "when to communicate". e.g., the binary action for "send / no-send": $a_i^C \in \mathcal{A}_i^C = \{0, 1\}$.

Denote $o_i^{C,t}$ as agent i's observation on the wireless environment at step t. We augment the observation space $\mathcal{O}$ as follows. Firstly, for each agent i, we concatenate the wireless and game observations, as $o_i^t = \left[o_i^{\mathcal{T},t} \| o_i^{C,t}\right]$. During communication, agents add $o_*^{C,t}$ to their original messages. The augmented observations help agents better understand both the wireless and game environments. Specifically, we augment the agents' observation by important network measurements (*e.g.*, radio signal strength, RSS), so that they can predict the dynamics due to *environment coupling*. Then we end-to-end train an intelligent communication scheme with neither pre-defined schedule nor prior knowledge on the wireless condition (unlike [11, 16]). As a result, agents learn "when to communicate" based on both the message relevance and the wireless channel conditions.

3 EXPERIMENTS

We let agents perform single-hop broadcast following [5, 8, 10, 11, 16]. We implement a 1-hop mobile network without Access Points (AP) as in [4]. We use "log distance path loss" model and model interference as receiver hearing multiple signals in range. We consider background noise and attenuation due to obstacles. The slotted p-CSMA protocol [6] is implemented for medium contention.

We evaluate both the on-policy and off-policy baselines. We also evaluate the corresponding variants for the proposed framework. For *off-policy* baselines, we evaluate the state-of-the-art value decomposition based algorithm, QMix [14]. Further, we additionally implement a communication-enhanced version of QMix by integrating the TarMAC design [2]. TarMAC is a state-of-the-art algorithm performing attention-based message aggregation. For *on-policy* baselines, we evaluate the state-of-the-art algorithms CommNet [16] and IC3Net [15], both of which are trained based on REINFORCE [19].

We conduct experiments on the standard MARL benchmark "Predator-Prey" with the following modifications: We introduce k obstacles that can affect both the game and wireless environments to better simulate realistic applications. Obstacles of different materials have different attenuation effect on the wireless signal passing it. We set gird size $n = 10$. There are 3 agents, each with vision $v = 0$, number of obstacles $k = 1$ and obstacle size $\ell = 9$.

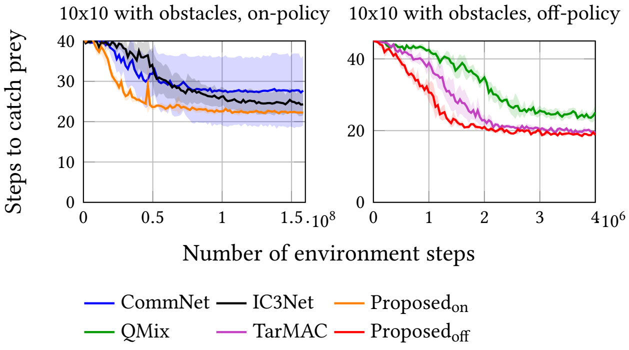


Figure 2: Comparison with state-of-the-art methods

We compare the proposed framework with state-of-the-art methods. We use RSS as the wireless observation. From Figure 2, we observe that for both the on-policy and off-policy algorithms, the proposed framework significantly shortens the number of steps to catch the prey. In addition, the variance of our curves are very small. For CommNet, the model hard-codes an all-to-all communication scheme: each step, all agents broadcast messages. In a realistic wireless network environment, sending more messages can reduce the number of successfully received messages due to increased chance of collision and interference. Therefore, it can be hard for the algorithm always performing broadcasting to stably learn a good policy – as reflected by the large variance of the CommNet curve. For IC3Net, the performance is better than CommNet since IC3Net has gated communication which mutes unimportant messages. Our framework further improves upon IC3Net due to its intelligent in both the "when" and "what" aspects covered in Section 2. As for the off-policy comparisons, QMix converges to a policy with significantly more steps. This shows the importance of communication. For our curve and TarMAC's, both converge to similar number of steps. However, ours converges faster and with smaller variance.

REFERENCES

[1] Micael S Couceiro, Rui P Rocha, and Nuno MF Ferreira. 2013. A PSO multi-robot exploration approach over unreliable MANETs. *Advanced Robotics* 27, 16 (2013), 1221–1234.

[2] Abhishek Das, Théophile Gervet, Joshua Romoff, Dhruv Batra, Devi Parikh, Mike Rabbat, and Joelle Pineau. 2019. Tarmac: Targeted multi-agent communication. In *International Conference on Machine Learning*. PMLR, 1538–1546.

[3] Yali Du, Bo Liu, Vincent Moens, Ziqi Liu, Zhicheng Ren, Jun Wang, Xu Chen, and Haifeng Zhang. 2021. Learning Correlated Communication Topology in Multi-Agent Reinforcement learning. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*. 456–464.

[4] Eduardo Feo Flushing, Michal Kudelski, Luca M Gambardella, and Gianni A Di Caro. 2014. Spatial prediction of wireless links and its application to the path control of mobile robots. In *Proceedings of the 9th IEEE International Symposium on Industrial Embedded Systems (SIES 2014)*. IEEE, 218–227.

[5] Jakob Foerster, Yannis M Assael, Nando de Freitas, and Shimon Whiteson. 2016. Learning to communicate with deep multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*. 2137–2145.

[6] Yi Gai, Shankar Ganesan, and Bhaskar Krishnamachari. 2011. The saturation throughput region of p-persistent CSMA. In *2011 Information Theory and Applications Workshop*. IEEE, 1–4.

[7] Ravi N Haksar and Mac Schwager. 2018. Distributed deep reinforcement learning for fighting forest fires with a network of aerial robots. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 1067–1074.

[8] Yedid Hoshen. 2017. Vain: Attentional multi-agent predictive modeling. *arXiv preprint arXiv:1706.06122* (2017).

[9] Datuk Prof Ir Ishak Ismail and Mohd Hairil Fitri Ja'afar. 2007. Mobile ad hoc network overview. In *2007 Asia-Pacific Conference on Applied Electromagnetics*. 1–8. https://doi.org/10.1109/APACE.2007.4603864

[10] Jiechuan Jiang and Zongqing Lu. 2018. Learning attentional communication for multi-agent cooperation. *arXiv preprint arXiv:1805.07733* (2018).

[11] Daewoo Kim, Sangwoo Moon, David Hostallero, Wan Ju Kang, Taeyoung Lee, Kyunghwan Son, and Yung Yi. 2019. Learning to Schedule Communication in Multi-agent Reinforcement Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SJxu5iR9KQ

[12] Yaru Niu, Rohan Paleja, and Matthew Gombolay. 2021. Multi-Agent Graph-Attention Communication and Teaming. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*. 964–973.

[13] Jorge Pena Queralta, Jussi Taipalmaa, Bilge Can Pullinen, Victor Kathan Sarker, Tuan Nguyen Gia, Hannu Tenhunen, Moncef Gabbouj, Jenni Raitoharju, and Tomi Westerlund. 2020. Collaborative multi-robot systems for search and rescue: Coordination and perception. *arXiv preprint arXiv:2008.12610* (2020).

[14] Tabish Rashid, Mikayel Samvelyan, Christian Schröder de Witt, Gregory Farquhar, Jakob N. Foerster, and Shimon Whiteson. 2018. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In *ICML*. 4292–4301.

[15] Amanpreet Singh, Tushar Jain, and Sainbayar Sukhbaatar. 2019. Individualized Controlled Continuous Communication Model for Multiagent Cooperative and Competitive Tasks. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rye7knCqK7

[16] Sainbayar Sukhbaatar, arthur szlam, and Rob Fergus. 2016. Learning Multiagent Communication with Backpropagation. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2016/file/55b1927fdafef39c48e5b73b5d61ea60-Paper.pdf

[17] Ming Tan. 1993. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*. 330–337.

[18] Rose E Wang, Michael Everett, and Jonathan P How. 2020. R-MADDPG for partially observable environments and limited communication. *arXiv preprint arXiv:2002.06684* (2020).

[19] Ronald J. Williams. 1992. Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning. *Mach. Learn.* 8, 3–4 (May 1992), 229–256. https://doi.org/10.1007/BF00992696

[20] Sai Qian Zhang, Qi Zhang, and Jieyu Lin. 2019. Efficient communication in multi-agent reinforcement learning via variance based control. *Advances in Neural Information Processing Systems* 32 (2019).