2022 IEEE 96th Vehicular Technology Conference (VTC2022-Fall) | 978-1-6654-5468-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/VTC2022-Fall57202.2022.10012942

Heterogeneous Mean-Field Multi-Agent Reinforcement Learning for Communication Routing Selection in SAGI-Net

Hengxi Zhang[1*], Huaze Tang[1*], Yuanquan Hu[1], Xiaoli Wei[1], Chenye Wu[2], Wenbo Ding[1,3], and Xiao-Ping Zhang[1,4]
[1] Tsinghua-Berkeley Shenzhen Institute, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China
[2] School of Science and Engineering, Chinese University of Hong Kong, Shenzhen, China
[3]RISC-V International Open Source Laboratory, Shenzhen, China
[4] Department of Electrical, Computer and Biomedical Engineering, Ryerson University, Toronto, Canada
{zhanghx20, thz21, huyq21}@mails.tsinghua.edu.cn, wei.xiaoli@sz.tsinghua.edu.cn, chenyewu@yeah.net, xzhang@ee.ryerson.ca
Corresponding author: Wenbo Ding, Email: ding.wenbo@sz.tsinghua.edu.cn

***Abstract*—The utilization of heterogeneous end devices such as the low earth orbit (LEO) satellite, unmanned aerial vehicles (UAVs) and ground users (GUs) deployed at different altitudes, known as the space-air-ground integrated network (SAGI-Net), can be quite promising towards a bunch of advanced applications. Whereas, the energy efficiency of the SAGI-Net communication system is a key criterion needed to be improved urgently in consideration that the inappropriate communication routing will undoubtedly cause a huge communication energy cost of the system especially with a large number of communication devices inside. In this paper, we proposed a novel communication routing selection model for the SAGI-Net system and established a heterogeneous multi-agent reinforcement learning (HMF-MARL) framework to optimize the communication energy efficiency of this system, where the mean-field theory was introduced to enhance the ability of classic MARL method while still maintaining a relatively low computational complexity. The experiment results show that the capacity of the heterogeneous multi-agent system has been improved by nearly 80% using the proposed HMF-MARL method compared with the classic MARL one, which hopefully shows the potential value on the implementation of the SAGI-Net system in the future.**

***Index Terms*—SAGI-Net, heterogeneous mean field, MARL, communication routing selection, computational complexity.**

I. Introduction

With the development of 5G technology, the Internet of things (IoT) plays a vital role in a myriad of applications and services nowadays, such as intelligent transportation systems, home automation, and smart factory [1]. To support these applications, huge computing demands emerge on IoT devices, which pose a challenge to current wireless communication networks as the current terrestrial communication paradigm cannot meet the requirement. As an extension to terrestrial communication, the space-air-ground integrated network (SAGI-Net) is proposed. SAGI-Net is a heterogeneous system that integrates satellites, air system consistent with unmanned aerial vehicles (UAV), and terrestrial communication system such as base station (BS) and aims to provide flexible network coverage and services [2].

SAGI-Net is viewed as a potential solution to meet the computing needs of IoT services and applications. However, there are still several challenging issues when employing SAGI-Net in IoT serves. Firstly, as the air system features high mobility, the communication routing in SAGI-Net faces dynamic channel conditions and coverage. In addition, different subsystems in SAGI-Net do not share the same communication interface and channels. Therefore, a thoughtful communication routing policy is required.

In recent years, reinforcement learning (RL) approaches, especially multi-agent reinforcement learning (MARL) has gained a surge of popularity in IoT network solutions since they hold efficacious promise to help address long-term decision-making problems in complex environments [3]. However, current MARL methods will face the curse of dimensionality when the agent number in the system goes large. Mean-field game (MFG) theory is an effective approach for handling such problems with a mass of agents or players, where the states or actions of all agents are established as two distributions or a joint distribution and each agent needs only to observe the distribution rather than every component [4, 5]. Therefore, the computational complexity of the numerical analysis for the multi-agent system can be apparently decreased. Chen *et al.* proposed a mean-field MARL method, named mean-field trust region policy optimization method, to obtain the optimal UAVs control [4]. To figure out the power control problem for ultra-dense device-to-device networks, Yang *et al.* formulated an MFG theoretic framework to acquire the energy and interference aware power control policy [5].

Nevertheless, the previous works on MFG mainly concentrate on homogeneous agents while there usually exist several or many categories of communication devices in a SAGI-Net, which leads to a heterogeneous communication system. Mondal

This work is supported in part by the grant from the Institute for Guo Qiang of Tsinghua University 2020GQG1004, by Tsinghua Shenzhen International Graduate School (QD2021013C QD2022009C), and also in part by Shenzhen Institute of Artificial Intelligence and Robotics for Society.

[*] These authors contribute equally.

Authorized licensed use limited to: Tsinghua University. Downloaded on November 01,2023 at 05:26:07 UTC from IEEE Xplore. Restrictions apply.

et al. discussed an integrated framework that combines the MFC theory and MARL from a theoretical point of view [6]. And some epidemic spreading models have also been established via heterogeneous mean-field approaches [7].

Considering that there are still few cross studies on the heterogeneous mean-field theory and MARL on SAGI-Net, we contribute to this work mainly from the following aspects: First of all, a specifically designed SAGI-Net communication routing model is proposed in a novel manner to optimize the transmission latency of the whole system, which is established as a distributionally robust optimization model. Second, we structure the SAGI-Net communication routing model as a Markov decision process of MARL. Finally, an enhanced MARL method, namely HMF-MARL, that combines heterogeneous mean-field theory with the classic MARL is implemented to obtain the numeral solutions.

II. System Model

We consider four communication categories at three different altitudes, with a bunch of GUs and a BS on the ground layer, the UAV swarm hovering in the air layer and a LEO satellite in the space layer, and five classes of communication links among these categories. The SAGI-Net communication system is presented in Fig. 1. The links started from GUs to BS, UAVs and satellite are modeled as GU-to-BS (G2B) link, GU-to-UAV (G2U) link and GU-to-Satellite (G2S) link, respectively. While each UAV can also establish two kinds of links from itself to BS and satellite, expressed similarly as the UAV-to-BS (U2B) link and UAV-to-Satellite (U2S) link.

At the beginning of each time slot, both GU and UAV swarm generate a series of data packets needed to be transmitted to the remote servers eventually. The BS and satellite are considered as the terminal transmission devices for each GU and UAV to transmit the data packets in the SAGI-Net, where the UAV in the air layer can be utilized as an intermediary for each GU in the ground to transmit the packet when the G2S or G2B link are temporally crowded or considering the transmit speed of them are relatively low. Specifically, GU and UAV swarm involve $N_1(=|\mathcal{N}_1|)$ and $N_2(=|\mathcal{N}_2|)$ homogeneous individuals respectively and can be further treated as a heterogeneous system $\mathcal{N}(=\{\mathcal{N}_1,\mathcal{N}_2\})$, where $\mathcal{N}_1=\{GU_1,GU_2,\dots,GU_{N_1}\}$ and $\mathcal{N}_2=\{UAV_1,UAV_2,\dots,UAV_{N_2}\}$. Each GU can select only one link from G2B, G2U, and G2S links to transmit data, while U2B or U2S are two links for each UAV to choose from. Note that the G2U link can only be generated if the UAV selected by the GU agrees to make this connection. Hereto, the objective of this SAGI-Net is to optimize the network capacity of the holistic system. In consideration that both GU groups and UAV swarm need to coordinate as a team in this scenario, we therefore model this SAGI-Net as a cooperative heterogeneous system.

The unicast protocol is employed in this work, where each GU or UAV can only select one objective to transmit its message, and the transmission between the same category, such as GU-to-GU and UAV-to-UAV, is prohibited to prevent message congestion. Since there exits three layers in this SAGI-Net, we allocate two bandwidths W_1 and W_2 that support the low-altitude and high-altitude transmission, respectively, in which low-altitude bandwidth W_1 supports G2B, G2U and U2B links while G2S and U2S links are on the high-altitude bandwidth W_2.

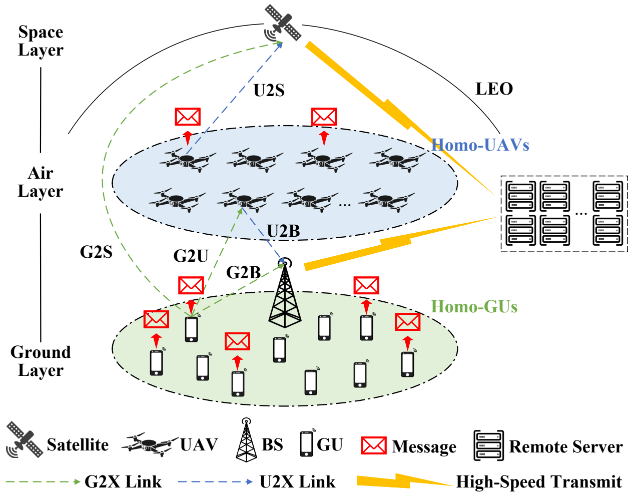

Fig. 1. A schematic illustration of SAGI-Net.

For the terrestrial Links including G2B, U2B and G2U links, in consideration of the interfering channels over the sub-band according to [3], the signal-to-interference-plus-noise ratio (SINR) of terrestrial link over h-th ($h\in[H]$) sub-band can be established as,

$$\gamma_i^a[h]=\frac{p_i^a[h]g_{i,B}[h]}{\sigma^2+\phi_i^{terra}[h]}, \tag{1}$$

and

$$\gamma_{i,j}^{G2U}[h]=\frac{p_i^{G2U}[h]g_{i,j}[h]}{\sigma^2+\phi_i^{terra}[h]}, \tag{2}$$

where $a\in\{G2B,U2B\}$, $p_i^a[h]$ and $g_{i,B}[h]$ refer to transmit power and channel of i-th ($i\in\mathcal{N}$) GU or UAV to the BS over h-th sub-band and $p_i^{G2U}[h]$ and $g_{i,j}[h]$ indicate the transmit power and the channel from the i-th GU to the j-th UAV over the h-th sub-band. σ^2 indicates the noise power. Meanwhile, considering that both GUs and UAVs keep transmitting data packets to the same BS and their transmissions will be hence interfered more by each other, we model the interference power $\phi_i^{terra}[h]$ as,

$$\begin{aligned}\phi_i^{terra}[h]=&\sum_{i'\in\mathcal{N},i'\neq i}x_{i',B}[h]p_{i'}^a[h]g_{i',B}[h]\\&+\sum_{i'\in\mathcal{N},i'\neq i}\sum_{j\in\mathcal{N}_2}x_{i',j}[h]p_{i',j}[h]g_{i',j}[h],\end{aligned} \tag{3}$$

where $x(\cdot)$ is the indicator function equal to 1 if the h-th sub-band is occupied and 0 otherwise. Note that if $i\in\mathcal{N}_2$, for all $i'\in\mathcal{N}_1$ there has that $i'\neq i$ and therefore, the inference from U2B links becomes $\sum_{i'\in\mathcal{N}_1}\sum_{j\in\mathcal{N}_2}x_{i',j}[h]p_{i',j}[h]g_{i',j}[h]$.

For non-terrestrial links including U2S and G2S links, since the satellite can be utilized to support long-range transmission due to its wide horizon, the G2S and U2S links in this SAGI-Net can be treated as a relatively smaller BS with relatively

Authorized licensed use limited to: Tsinghua University. Downloaded on November 01,2023 at 05:26:07 UTC from IEEE Xplore. Restrictions apply.

wide bandwidth for receiving a series of messages from both GUs and UAVs. Therefore, we express the SINR of G2S and U2S link over h-th ($h \in [H]$) sub-band as,

$$\gamma_i^b[h] = \frac{p_i^b[h] g_{i,S}[h]}{\sigma^2 + \phi_i^b[h]}, \tag{4}$$

where $b \in \{G2S, U2S\}$, $p_i^b[h]$ and $g_{i,S}[h]$ indicate the transmit power and the interfering channel from the i-th GU or UAV to the satellite over the h-th sub-band. And considering there are only G2S and U2S links on the high-altitude bandwidth W_2, we use

$$\phi_i^b[h] = \sum_{i' \in \mathcal{N}, i' \neq i} x_{i',S}[h] p_{i'}^b[h] g_{i',S}[h], \tag{5}$$

to indicate the interference power of i'-th non-terrestrial links over h-th sub-band.

Furthermore, based on all types of SINRs above, the transmit rates of different links over the same h-th sub-band can be hence established as

$$R^c[h] = W_d \log(1 + \gamma^c[h]), \tag{6}$$

where $c \in \{G2B, G2S, G2U, U2B, U2S\}$ represents different types of links and $d \in \{1, 2\}$ indicates the bandwidth occupied for transmission.

The latency of transmitting the data packet or message between communication nodes depends mainly on and packet size B and the transmit rate over the specific sub-band, given as,

$$L^c[h] = \frac{B}{R^c[h]}, \tag{7}$$

where the size of message B is considered a fixed constant for both GUs and UAVs in this work.

According to the distributionally robust optimization model, the worst-case, i.e. the maximum latency, in the communication system determines the quality of the transmission [8, 9]. With the maximum latency over a series of fixed time slots ΔT, the objective in this SAGI-Net is to minimize the expected maximum latency over a series of fixed time slots ΔT, mathematically formulated as,

$$\begin{aligned}
&\min_{\mathbf{X},\mathbf{P},\mathbf{H}} \max_{\omega \in \Omega} \mathbb{E}_{\omega,\Delta T}\left[L^c(\mathbf{X},\mathbf{P},\mathbf{H})\right] \\
&\text{s.t.}(a) \sum_j x_{i,j} + x_{i,B} + x_{i,S} = 1, \forall GU_i, i \in \mathcal{N}_1, \\
&\quad (b)\ x_{j,B} + x_{j,S} = 1, \forall\ UAV_j, j \in \mathcal{N}_2, \\
&\quad (c)\ x_{i,j}, x_{i,B}, x_{i,S}, x_{j,B}, x_{j,S} \in \{0,1\}, \\
&\quad (d)\ p_{i,j}, p_{i,B}, p_{i,S}, p_{j,B}, p_{j,S} \geq 0,
\end{aligned} \tag{8}$$

where ω denotes the worst-case distribution realization in all case realization Ω.

III. Heterogeneous Mean-Field MARL in SAGI-NET

In this section, we have specifically formulated the distributionally robust optimization model of SAGI-Net into a MARL framework in the first step. Then the classic MARL algorithm is utilized to cope with this optimization problem. Afterwards, we further integrate the heterogeneous mean-field theory into the classic MARL to enhance its performance.

A. Markov Decision Process of MARL

RL is a goal-oriented learning approach in which an agent learns to achieve the optimal long-term goal through trail and error in a specific environment. The agent is rewarded when its behaviour leads to satisfactory outcomes, or get punished when its behaviour leads to bad results. A typical problem in RL can be modeled as a Markov Decision Process (MDP) which is composed of a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, R, \gamma \rangle$, where $\mathcal{S}$ denotes the set of possible states of the environments, $\mathcal{A}$ denotes the set of the agent's possible actions, $\mathcal{P}$ denotes the transition distribution $P(s'|s,a)$, R denotes the reward function evaluating a transition from s to s' as a result of action a and $\gamma \in [0,1)$ is the discounted factor that represents the value of time. The goal of RL is to learn a policy $\pi : \mathcal{S} \to \mathcal{A}$ to maximize the expected cumulative reward. Based on this we can define the state-action function (Q-function) for a policy π as:

$$Q^\pi(s,a) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t R_t | s_0 = s, a_0 = a \right]. \tag{9}$$

MARL is an extended form of RL where multiple agents interact in the environment. Similarly, a MARL problem can be represented as a tuple $\langle \mathcal{N}, \mathcal{S}, \boldsymbol{\mathcal{A}}, \mathcal{P}, R, \gamma \rangle$, where $\mathcal{N} = \{1, \cdots, n\}$ is the set of agents, $\mathcal{S}$ denotes the global state space and $\boldsymbol{\mathcal{A}} = \Pi_{j=1}^n \mathcal{A}^j$ is the joint action space for n agents. The remaining components have the same meaning as in the MDP of single-agent RL. In MARL, the Q-function for each agent j can be reformulated as:

$$Q^\pi(s_j, a_j, s_{-j}, a_{-j}) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t R_i^t | s, a, s_{-j}, a_{-j} \right], \tag{10}$$

where s_{-j} and a_{-j} are the joint state and joint action of all agents except agent j.

B. Heterogeneous Mean-Field Cooperative MARL

In this part, we formulate our problem into a heterogeneous cooperative MARL. We have two heterogeneous sets of agents $\mathcal{N}_1$ and $\mathcal{N}_2$, which represent the population of UAVs and GUs, respectively. And we assume the agents in each heterogeneous set are homogeneous, i.e., they share the same state sapce $\mathcal{S}_l$ and action space $\mathcal{A}_l$ for class l. Hence our global state space and joint action space can be represented as $\boldsymbol{\mathcal{S}} = (\mathcal{S}_1)^{N_1} \times (\mathcal{S}_2)^{N_2}$ and $\boldsymbol{\mathcal{A}} = (\mathcal{A}_1)^{N_1} \times (\mathcal{A}_2)^{N_2}$. In the cooperative setting, the overall reward of the system is the summation of the rewards of each individual agent. We assume that the Q-function can be decomposed as the pairwise local interactions:

$$Q^j(\boldsymbol{s},\boldsymbol{a}) = \frac{1}{N_1^j} \sum_{k_1} Q^j(\boldsymbol{s}, a^j, a_1^{k_1}) + \frac{1}{N_2^j} \sum_{k_2} Q^j(\boldsymbol{s}, a^j, a_2^{k_2}), \tag{11}$$

where $k_1 \in \mathcal{N}_1(j)$ and $k_2 \in \mathcal{N}_2(j)$ are the index sets of the neighboring UAV agents and GU agents with respect to agent j with size $N_1^j = |\mathcal{N}_1(j)|$ and $N_2^j = |\mathcal{N}_2(j)|$ [1]. And this formulation can be rewritten as

$$Q^j(\boldsymbol{s},\boldsymbol{a}) = \frac{1}{N^j} \sum_{k=1}^{N^j} Q^j(\boldsymbol{s}, a^j, a_1^k, a_2^k), \tag{12}$$

[1] We use j as the index of agents for the whole heterogeneous system.

Authorized licensed use limited to: Tsinghua University. Downloaded on November 01,2023 at 05:26:07 UTC from IEEE Xplore. Restrictions apply.

where $N^j = N_1^j + N_2^j$ and we add a placeholder agent of the other class in each term of (11).

Since we have an enormous amount of agents in each heterogeneous subsystem, conventional control methods can be impractical due to the curse of dimensionality of the multi-agent problem. Here we use the idea of mean field theory [10] to tackle the scalability issue of MARL in the SAGI-Net system. Using mean-field approximation [10, 11], the pairwise interactions $Q^j(\boldsymbol{s}, a^j, a_1^k, a_2^k)$ can be expended by Taylor's theorem and expressed as the mean effect of the neighboring agents:

$$Q^j(\boldsymbol{s},\boldsymbol{a}) = \frac{1}{N^j}\sum_{k=1}^{N^j} Q^j(\boldsymbol{s},a^j,a_1^k,a_2^k) \approx Q^j(\boldsymbol{s},a^j,\overline{a}_1^j,\overline{a}_2^j), \quad (13)$$

where $\overline{a}_l^j (l \in \{1,2\})$ denotes the mean action of agents belonging to class l in the neighborhood of agent j.

C. Implementation

We use deep Q-learning to learn the state-action function in heterogeneous mean-field multi-agent reinforcement learning (HMF-MARL). We roll out current policy to collect experiences $\langle \boldsymbol{s}, \boldsymbol{a}, \boldsymbol{r}, \boldsymbol{s'}, \overline{\boldsymbol{a}}_1, \overline{\boldsymbol{a}}_{\boldsymbol{2}} \rangle$ and store them to a replay buffer. The model parameter of Q network is learned by sampling batches of D transitions from the replay buffer and minimizing the squared temporal difference error:

$$\mathcal{L}(\theta^j) = \sum_{d=1}^{D}\left[\left(y_d^j - Q^j(\boldsymbol{s},a^j,\overline{a}_1^j,\overline{a}_2^j;\theta^j)\right)^2\right], \quad (14)$$

where $y^j = r^j + \gamma \max_{a^{j'}} Q^j(\boldsymbol{s'}, a^{j'}, \overline{a}_1^{j'}, \overline{a}_2^{j'}; \theta_-^j)$. θ_-^j are the parameters of agent j's target network that are periodically updated by θ^j. The mean actions $\overline{a}_1^{j'}$ and $\overline{a}_2^{j'}$ for next state are predicted by the current policy.

IV. Experiments and Results

For implementing the HMF-MARL method in the SAGI-Net, we have specifically designed a SAGI communication scenario. Since there does not exist an overall channel model for SAGI, we combine different channel models to characterize different channels in SAGI-Net. Specifically, we model the G2B channel as WINNER II channel model [12], U2B channel following the definition in 3GPP TR 36.777 Rel. 15 [13], G2U channel following definition in [14] and non-terrestrial links (G2S and U2S) following the definition in 3GPP TR 38.811 Rel. 15 [15]. The simulation experiment settings are listed as Table. I.

To evaluate the performance of the proposed HMF-MARL approach, we then set the classic MARL as the comparison, which is also specifically established based on this SAGI-Net communication scenario. And the random experiment, where both GUs and UAVs just select actions randomly, is treated as the lower bound. The rewards with different methods in the training phase are presented in Fig. 2.

Fig. 2 shows that even though the performance of classic MARL is about 1.5 times the lower bound, it is still relatively inferior due to the limited information sharing for this SAGI-Net multi-agent system. While the performance of proposed HMF-MARL method has dramatically reaches nearly three times the lower bound, where both each GU and UAV not only keep collecting the full observation over the whole communication networks, but utilize the mean field of action from both agent classes (i.e., GUs and UAVs) as the critical factors for making decision.

In addition, for further investigating the effectiveness and stability of this HMF-MARL method, the package sizes for both GUs and UAVs to transmit are set as $[10, 20, 30, 40, 50]$ Kb, respectively. And the accumulated rewards obtained by the multi-agent system using HMF-MARL are shown in Fig. 3.

We show in Fig. 3 that the accumulated rewards in convergence phase (after about 2500 episodes) decrease with the

TABLE I
Simulation Settings [12–15]

Network Parameter	**Detail**
Number of GU	5
Number of UAV	3
Number of BS	1
Number of Satellite	1
Carrier Frequency of terrestrial links	2 GHz
Bandwidth of terrestrial links	10 MHz
Carrier Frequency of non-terrestrial links	30 GHz
Bandwidth of non-terrestrial links	50 MHz
GU & UAV Transmit Power	[23,10,5,-100] dBm
Number of sub-band	10
BS Antenna Gain	8 dBi
UAV & GU Antenna Gain	3 dBi
GU & UAV Receiver Noise Figure	9 dB
BS Receiver Noise Figure	5 dB
Noise Power	-114 dBm
Time Slot for Package Delivery	100 ms
Message Size	$[10, 20, \ldots, 50] \times 1060$ bytes
DQN Parameter	**Detail**
Number of Episode	3000
Time Step	100
Learning Rate	0.001
Discount Factor	0.99
Gradient Descent Frequency	100
Target Q-Network Update Frequency	2000
Batch Size	256
Node Activation DQN	ReLU Function

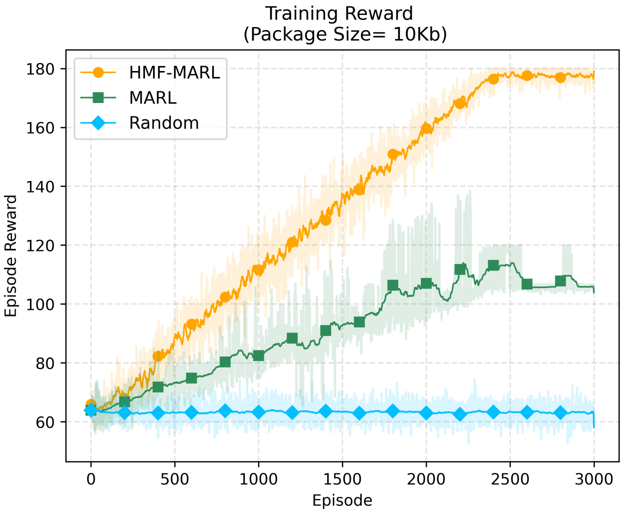

Fig. 2. The episode rewards with increasing training iterations.

Authorized licensed use limited to: Tsinghua University. Downloaded on November 01,2023 at 05:26:07 UTC from IEEE Xplore. Restrictions apply.

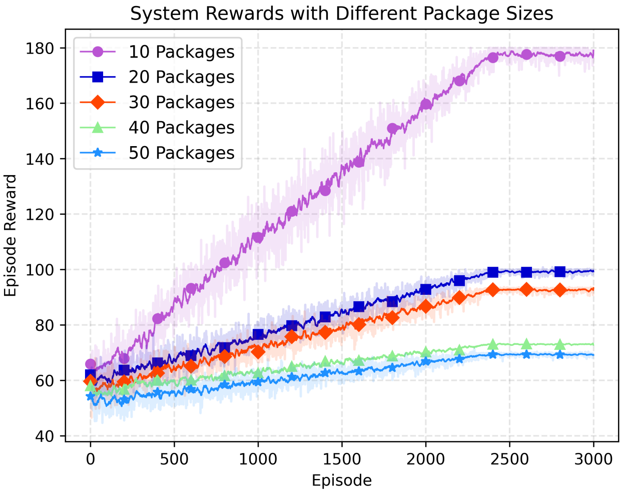

Fig. 3. Episode rewards of HMF-MARL with different package sizes.

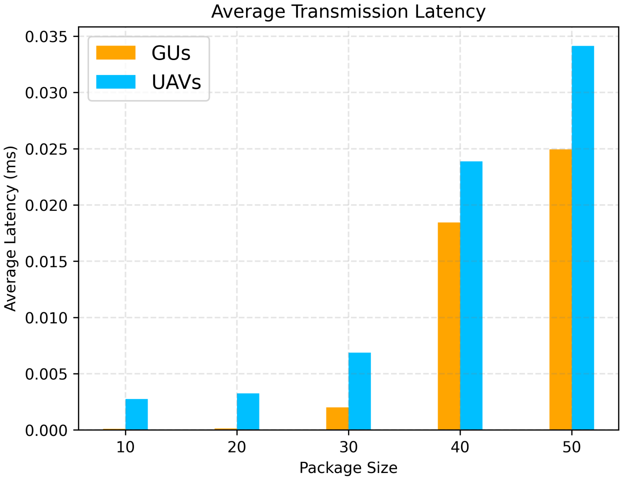

Fig. 4. The average latency of GUs and UAVs in HMF-MARL protocol.

increment of the package sizes. Since all the communication resources, such as bandwidth, spectrum, transmit power, etc., for each agent of different classes to transmit do not change while the payload size rises, it becomes more difficult for GUs and UAVs to complete the package delivery in the limited time slots, which results in the decline of the accumulated rewards because the reward is strictly bound to the objective function, i.e. (8). Meanwhile, the stably and promptly increasing rewards in HMF-MARL scenario indicate the effectiveness of this algorithm.

Furthermore, we investigate the average transmission latency of GUs and UAVs to study the transmission capacity of the whole SAGI-Net communication networks. Fig. 4 presents that the average transmission latency of GUs and UAVs both rise with the increment of the package size due to the larger payload size and constant and limited communication resources. And the GU class generally takes less time to transmit the data package in contrast to the UAV class since each GU has more communication routing options than the UAV does, which makes the routing selection for the GU class more flexible.

V. Conclusion

In this work, we have proposed a specifically designed HMF-MARL approach for tackling the communication routing selection problem over the SAGI-Net scenario in a novel manner. Considering that there are multiple types of communication nodes and different communication protocols are correspondingly required in this complex network system, we choose to introduce a novel heterogeneous mean-field theory and integrate it into the classic MARL approach for enhancing its performance. The experiment results show that the capacity of the heterogeneous multi-agent system has been improved by nearly 80% using the proposed HMF-MARL method compared with the classic MARL one, which may provide a promising way towards implementing the heterogeneous and distributed MARL protocol in the SAGI-Net communication networks.

References

[1] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of things: A survey on enabling technologies, protocols, and applications," *IEEE Communications Surveys Tutorials*, vol. 17, no. 4, pp. 2347–2376, 2015.

[2] J. Liu, Y. Shi, Z. M. Fadlullah, and N. Kato, "Space-air-ground integrated network: A survey," *IEEE Communications Surveys Tutorials*, vol. 20, no. 4, pp. 2714–2741, 2018.

[3] L. Liang, H. Ye, and G. Y. Li, "Spectrum sharing in vehicular networks based on multi-agent reinforcement learning," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2282–2292, 2019.

[4] D. Chen, Q. Qi, Z. Zhuang, J. Wang, J. Liao, and Z. Han, "Mean field deep reinforcement learning for fair and efficient uav control," *IEEE Internet of Things Journal*, vol. 8, no. 2, pp. 813–828, 2020.

[5] C. Yang, J. Li, P. Semasinghe, E. Hossain, S. M. Perlaza, and Z. Han, "Distributed interference and energy-aware power control for ultra-dense d2d networks: A mean field game," *IEEE Transactions on Wireless Communications*, vol. 16, no. 2, pp. 1205–1217, 2016.

[6] W. U. Mondal, M. Agarwal, V. Aggarwal, and S. V. Ukkusuri, "On the approximation of cooperative heterogeneous multi-agent reinforcement learning (marl) using mean field control (mfc)," *arXiv preprint arXiv:2109.04024*, 2021.

[7] C. Li, R. van de Bovenkamp, and P. Van Mieghem, "Susceptible-infected-susceptible model: A comparison of n-intertwined and heterogeneous mean-field approximations," *Physical Review E*, vol. 86, no. 2, p. 026116, 2012.

[8] E. Delage and Y. Ye, "Distributionally robust optimization under moment uncertainty with application to data-driven problems," *Operations research*, vol. 58, no. 3, pp. 595–612, 2010.

[9] Y. Chen, B. Ai, Y. Niu, H. Zhang, and Z. Han, "Energy-constrained computation offloading in space-air-ground integrated networks using distributionally robust optimization," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 11, pp. 12 113–12 125, 2021.

[10] Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, and J. Wang, "Mean field multi-agent reinforcement learning," in *International conference on machine learning*. PMLR, 2018, pp. 5571–5580.

[11] S. G. Subramanian, P. Poupart, M. E. Taylor, and N. Hegde, "Multi type mean field reinforcement learning," *arXiv preprint arXiv:2002.02513*, 2020.

[12] M. Döttling, W. Mohr, and A. Osseiran, *WINNER II Channel Models*, 2010, pp. 39–92.

[13] *3rd Generation Partnership Project; Technical Specification Group Radio Access Network;Study on Enhanced LTE Support for Aerial Vehicles: (Release 15)*, Standard 3GPP TR, Dec 2017.

[14] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal lap altitude for maximum coverage," *IEEE Wireless Communications Letters*, vol. 3, no. 6, pp. 569–572, 2014.

[15] *3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on New Radio (NR) to support non-terrestrial networks : (Release 15)*, Standard 3GPP TR, Sep 2020.

Authorized licensed use limited to: Tsinghua University. Downloaded on November 01,2023 at 05:26:07 UTC from IEEE Xplore. Restrictions apply.