Foundation Model-aided Multi-agent Reinforcement Learning for Random Access Network Optimization

Myeung Suk Oh
Department of Electrical and Computer Engineering
The Ohio State University
Columbus, OH 43210
oh.746@osu.edu

Alvaro Velasquez
Department of Computer Science
University of Colorado Boulder
Boulder, CO 80309
alvaro.velasquez@colorado.edu

Jia Liu
Department of Electrical and Computer Engineering
The Ohio State University
Columbus, OH 43210
liu.1736@osu.edu

Abstract

Random access (RA) is one of the most foundational medium access control (MAC) layer scheduling schemes for handling unpredictable data traffic from multiple terminals and serves as the basis for modern carrier-sense multiple access (CSMA) protocols. While multi-agent reinforcement learning (MARL) has been explored to optimize RA-based networks, its reliance on experience-driven, distributed policy learning incurs significant training overhead for each optimization task, limiting their feasibility in real-world applications. In this work, we propose to leverage a foundation model (FM) to improve MARL efficiency across diverse RA network optimization tasks. Specifically, we design an FM-aided actor-critic algorithm within a consensus-based decentralized MARL architecture and provide its convergence analysis. Numerical evaluations show that our proposed method enhances MARL efficiency for RA network optimization.

1 Introduction

1) Background and Motivations Random access (RA) techniques have long been woven into the fabric of modern wireless networks at the medium access control (MAC) layer, with prominent examples including CSMA/CA in Wi-Fi and LTE-LAA [1]. In RA networks, multiple devices share a common channel and make independent transmission decisions without centralized scheduling. This decentralized design enables scalable and flexible operation in a wide range of applications, such as the Internet of Things (IoT) [2], machine-type communications (MTC) [3], and smart grid infrastructure [4], where many devices are connected with sporadic traffic patterns. A central challenge in RA lies in optimizing network performance while mitigating collisions among devices. Despite decades of research, existing solutions often rely on heuristic methods that fail to achieve strong performance or idealized analytical models that inadequately capture the complexities of real-world environments.

2) Multi-agent Reinforcement Learning Approaches and Foundation Models Recent advances in machine learning (ML) have allowed intelligent strategies for optimizing RA networks (e.g., [5, 6, 7]). Given their decentralized nature, RA networks can be modeled as distributed decision-making

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI and ML for Next-Generation Wireless Communications and Networking (AI4NextG).

systems, making them well-suited to multi-agent reinforcement learning (MARL) framework. This has motivated many MARL-based approaches for RA optimization [8, 9, 10, 11, 12, 13, 14] (see Appendix A.1 for details). However, conventional MARL methods often suffer from task-specific overfitting as they are primarily tailored to a single objective (e.g., minimizing collisions). Achieving high performance in different objectives often demands complex models and extensive re-training, which reduces efficiency and limits generalization across tasks. These limitations highlight the need for more efficient MARL frameworks for RA optimization.

In recent years, foundation models (FMs), which are pretrained on large datasets and finetuned for downstream tasks, have achieved astonishing successes in natural language processing and emerged as powerful tools for enhancing reinforcement learning (RL). Self-supervised FMs can capture data dynamics as sequence models (e.g., MAMBA [15]) and provide reward-agnostic representations that accelerate online learning. With FMs also gaining attention in wireless communications recently (see Appendix A.2 for details), integrating them into MARL presents a promising direction for efficient RA network optimization. However, a key challenge lies in the deployment of FMs within existing RL architectures (e.g., actor-critic) and designing MARL algorithms that fully leverage their representational power.

3) Our Approach and Contributions We propose an FM-aided MARL framework for optimizing RA networks under a fully decentralized architecture, where policy learning proceeds *without* centralized training. To ensure global convergence, we employ an average consensus mechanism [16], allowing devices to exchange local information with neighbors. Our design builds on the actor-critic framework (e.g., A2C [17], SAC [18], DDPG [19], PPO [20]), where an FM serves as a reward-agnostic backbone. A reward-specific head is then attached to form the critic network. The FM is pretrained in a self-supervised manner on reward-independent network data, capturing system dynamics and generating expressive latent representations. This FM-aided actor-critic accelerates online policy learning while maintaining comparable or superior performance to end-to-end training.

Our main contributions are as follows: **1)** We introduce a new method to incorporate an FM as a reward-agnostic critic backbone inspired by the SMART architecture [21] pretrained for forward dynamics prediction; **2)** We define Markov decision process (MDP) parameters for three RA optimization objectives: fair age of information (AoI), maximum sum-rate, and fair rate. We use distinct locally computable reward functions compatible with decentralized MARL; **3)** We provide a mathematical analysis of the convergence properties of our decentralized FM-aided actor–critic algorithm; and **4)** We conduct numerical experiments to demonstrate consistent improvements in learning speed and performance across diverse RA optimization tasks.

2 System Model

1) Network Configuration: We consider an RA network with K single-antenna devices attempting to transmit data packets to an N_{r}-antenna access point (AP). The network operates in slotted time over a finite horizon T, with time slots indexed by $t \in \{1, 2, \ldots, T\}$. Following the CSMA/CA protocol, each device $k \in \mathcal{K} = \{1, 2, \ldots, K\}$ adopts a listen-before-talk (LBT) mechanism, performing clear channel assessment (CCA) prior to transmission. If the channel becomes idle, the device decides whether to transmit a packet to the AP over τ time slots. A successful transmission (i.e., no collision) triggers an acknowledgment (ACK) from the AP, allowing the device to proceed with next packet. In case of collision, no ACK is sent, and the device waits for the next opportunity to retransmit. We assume a saturated traffic condition, i.e., all devices always have packets to transmit. To support decentralized operation, no central controller is assumed. Instead, devices exchange local information with neighbors defined by an undirected graph $\mathcal{G} = (\mathcal{K}, \mathcal{E})$, where $\mathcal{E}$ represents connectivity [16].

2) Data Transmission Model: The single-input multiple-output (SIMO) channel between device k and the AP at time slot t is given by

$$\boldsymbol{g}_{k,t} = \sqrt{\eta_{k,t}}\boldsymbol{h}_{k,t}, \tag{1}$$

where $\boldsymbol{h}_{k,t} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_{N_{\mathsf{r}}})$ represents small-scale fading and $\eta_{k,t}$ denotes the distance-dependent large-scale fading factor. Then, the achievable data rate of device k at time slot t can be obtained by

the Shannon capacity formula:

$$R_{k,t} = B \log_2 \Big(1 + \frac{p_k \|\boldsymbol{g}_{k,t}\|_2^2}{BN_0}\Big), \tag{2}$$

where B is the bandwidth, p_k is the transmit power of device k, and N_0 is the noise spectral density [22]. Assuming a block-fading channel (i.e., the channel remains static during each packet transmission), the amount of data which device k can transmit at time slot t is expressed as $m_{k,t} = R_{k,t} T_{\mathsf{s}} \tau$, where T_{s} is the slot duration.

3) Downstream Tasks: We consider the following three objectives for our downstream tasks:

i) Fair-AoI (Worst-case AoI Minimization): AoI of device k can be quantified as the number of time slots elapsed since its last successful packet transmission, which we denote as $\ell_{k,t}$. Let $\mathcal{T}_{k,t} \subseteq \{1, 2, \ldots, t\}$ denote the set of time slot indices at which device k has successfully transmitted a packet during the time horizon $t \leq T$. We minimize the worst-case AoI among all devices using

$$\min \max_{k \in \mathcal{K}, t \in \mathcal{T}_{k,T}} \ell_{k,t}. \tag{3}$$

This task is useful in applications where timely information updates are important.

ii) Max-rate (Sum-rate Maximization with Minimum Rate Guarantees): We define

$$x_{k,t} = \frac{1}{tT_{\mathsf{s}}} \sum_{t' \in \mathcal{T}_{k,t}} m_{k,t'} \tag{4}$$

as the throughput of device k over the time horizon $t \leq T$. To maximize sum-rate, i.e., $\sum_{k \in \mathcal{K}} x_{k,T}$, it is ideal to prioritize devices with higher data rates, but this can result in poor fairness. To prevent this, we impose a minimum rate requirement R_{min} and formulate our objective as

$$\begin{aligned} &\max \sum_{k \in \mathcal{K}} x_{k,T} \\ &\text{s.t. } x_{k,T} \geq R_{\mathsf{min}}, \ \forall k \in \mathcal{K}. \end{aligned} \tag{5}$$

This problem seeks to maximize the sum-rate while ensuring each device's rate to be at least R_{min}.

iii) Fair-rate (Worst-case Rate Maximization): To ensure balanced per-device rate, we maximize the minimum throughput across all devices over the time horizon T, which can be expressed as

$$\max \min_{k \in \mathcal{K}} x_{k,T}. \tag{6}$$

This formulation ensures that no device is significantly disadvantaged in terms of data rate.

3 Foundation Model-aided Multi-agent Reinforcement Learning Framework

1) MARL Problem Formulation: We formulate the FM-aided MARL problem by the tuple $\{\mathcal{S}, \{\mathcal{A}_k\}_{k \in \mathcal{K}}, P, \{R_k\}_{k \in \mathcal{K}}\}$, where $\mathcal{S}$ is the global state space, $\mathcal{A}_k$ is the action set for device k, P is the state transition probability, and R_k is the local reward function for device k. At a time slot t, each device k at a global state $s_t \in \mathcal{S}$ selects an action $a_{k,t} \in \mathcal{A}_k$ according to its local policy π_k, i.e., $a_{k,t} \sim \pi_k(\cdot|s_t)$. As the state transitions from s_t to s_{t+1}, each device k receives a reward $r_{k,t} = R_k(s_t, \boldsymbol{a}_t)$, where $\boldsymbol{a}_t = [a_{1,t}, a_{2,t}, \ldots, a_{K,t}]$ is the joint action vector.

We define our state as

$$s_t = \big(\{\widehat{\xi}_{k,t}\}_{k \in \mathcal{K}}, \{\widehat{\ell}_{k,t}\}_{k \in \mathcal{K}}, q_t\big), \tag{7}$$

where $\widehat{\xi}_{k,t} = \frac{\xi_{k,t} - \min_k \mathbb{E}_t[\xi_{k,t}]}{\max_k \mathbb{E}_t[\xi_{k,t}] - \min_k \mathbb{E}_t[\xi_{k,t}]}$ is the min-max normalized signal-to-noise ratio (SNR) for device k, $\widehat{\ell}_{k,t} = \ell_{k,t}/\lambda$ is the normalized AoI with λ as a scaling factor, and $q_t \in \{0, 1\}$ is the binary channel occupancy indicator. We use dB scale for SNR, i.e., $\xi_{k,t} = 10 \log_{10} \frac{p_k \|\boldsymbol{g}_{k,t}\|_2^2}{BN_0}$. For each device k, $\widehat{\xi}_{k,t}$ and $\widehat{\ell}_{k,t}$ are local information. Following [10, 13, 14], we consider AoIs of other devices $\{\widehat{\ell}_{k',t}\}_{k' \in \mathcal{K} \setminus k}$ and q_t to be observable from the broadcast ACK packets and CCA phase,

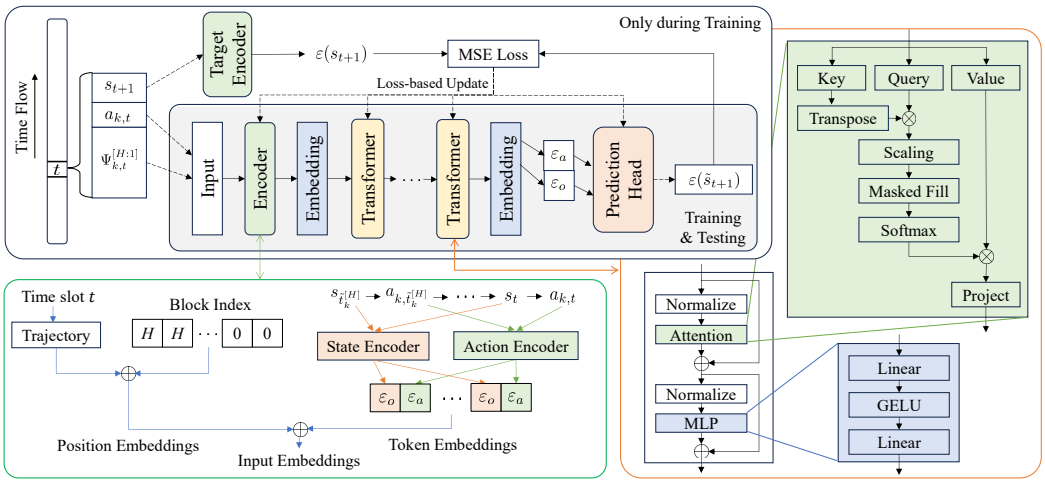


Figure 1: The architecture of FM as a reward-agnostic critic backbone. The model takes an state-action history and extracts embeddings via state and action encoders. Transformer blocks are used to process the embeddings. The output of a prediction head is sent to the reward-specific critic head.

respectively. We assume that SNRs of other devices $\{\widehat{\xi}_{k',t}\}_{k' \in \mathcal{K} \setminus k}$ are available via local information exchange. We consider a discrete action $a_{k,t} \in \mathcal{A}_k = \{0, 1\}$, where 0 indicates *wait* and 1 represents *transmit*.

We now define local reward functions corresponding to each downstream objective. First, the local reward for fair-AoI task is defined as

$$r_{k,t}^{\mathsf{FA}} = -\ell_{k,t}/(\omega_{\mathsf{FA}} K), \tag{8}$$

where $\omega_{\mathsf{FA}} K$ is a scaling factor. Since $\ell_{k,t} = 0$ only upon successful transmission, this reward encourages devices to avoid collisions and prioritizes those with higher AoI, thus promoting fairness.

Next, the local reward for max-rate task is given by

$$r_{k,t}^{\mathsf{MR}} = x_{k,t}/\omega_{\mathsf{MR}} + D(x_{k,t}), \tag{9}$$

where ω_{MR} is a scaling parameter, and $D(x_{k,t})$ is a penalty function defined as

$$D(x_{k,t}) = \begin{cases} \log\left(\frac{x_{k,t}+\omega_{\mathsf{P1}} R_{\min}}{(1+\omega_{\mathsf{P1}}) R_{\min}}\right)^{\omega_{\mathsf{P2}}/K}, & \text{if } x_{k,t} \le R_{\min}, \\ 0, & \text{otherwise}, \end{cases} \tag{10}$$

with scaling factors ω_{P1} and ω_{P2}. Note that (10) is bounded in $[\log(\omega_{\mathsf{P}}/(1+\omega_{\mathsf{P}})), 0]$. This reward derives each device to maintain the minimum rate requirement while maximizing the sum-rate.

Lastly, we adopt the α-fairness utility function [10] and define the reward for fair-rate task as

$$r_{k,t}^{\mathsf{FR}} = \frac{1}{1-\alpha}\left(\frac{x_{k,t}}{\omega_{\mathsf{FR1}}/K} + \omega_{\mathsf{FR2}}\right)^{1-\alpha}, \tag{11}$$

where $\alpha \in (1, \infty)$ controls the level of fairness, and ω_{FR1} and ω_{FR2} are scaling factors.

Note that, to reflect the condition of fully decentralized MARL, these reward functions are restricted to be local. For the consensus step, we let $r_{k,t} \in \{r_{k,t}^{\mathsf{FA}}, r_{k,t}^{\mathsf{MR}}, r_{k,t}^{\mathsf{FR}}\}$ to be exchanged across the devices through local communication links defined by $\mathcal{G}$.

2) Foundation Model (FM) Architecture: We provide the details of our FM, which supports the actor-critic framework (see Appendix B for details) as a reward-agnostic critic backbone. The FM is pretrained in a *self-supervised* manner and thus does *not* require reward labels. This allows the FM to be reward-agnostic and transferrable across different RL tasks. Our FM pretraining is inspired by the SMART framework [21], which uses a control transformer architecture to facilitate

self-supervised learning across different pretraining goals. Similar to SMART, we employ encoders to extract meaningful embeddings and utilizes the attention mechanism within transformers to perform a prediction task. In this paper, we specifically adopt the *forward dynamics prediction* task for self-supervised pretraining, which aligns naturally with the core principle of RL to optimize future outcome predictions.

The architecture of our FM is shown in Fig. 1, which consists of three main parts: encoders, transformers, and a prediction head. The encoders first map the input into a sequence of token embeddings, which are augmented with position embeddings to convey temporal information. The combined embeddings are then passed through a multi-transformer module, where each block consists of a self-attention layer and a multi-layer perceptron (MLP). The output from the transformer module is fed into the prediction head, which generates an embedding that represents next-step state. During the pretraining, the FM is optimized such that the loss between the predicted and actual next-state embeddings is minimized. By integrating the self-supervised FM, we allow our MARL agents to utilize effective representations of the environment dynamics during their online learning.

Similar to [10, 13, 14], we assume that each device can store the H latest states and actions. Let $\tilde{t}_k^{[h]}$ denote the time slot when the h-th latest action was taken by device k. Then, the state-action history at time slot t can be defined as

$$\Psi_{k,t}^{[H:1]} = \{s_{\tilde{t}_k^{[H]}}, a_{k,\tilde{t}_k^{[H]}}, \ldots, s_{\tilde{t}_k^{[1]}}, a_{k,\tilde{t}_k^{[1]}}, s_t\}, \tag{12}$$

which we use as an input to both actor and critic. This enables the model to better reflect the temporal dynamics of RA environment.

3) Decentralized MARL Algorithm Implementation: Fig. 2 shows the overall framework of our FM-aided decentralized MARL for RA network optimization. Each device $k \in \mathcal{K}$ maintains a history buffer to store past states and actions and updates its own actor and critic models. Connected devices can exchange local information (i.e., rewards and SNRs) with one another. The critic network consists of two parts: a reward-agnostic FM and a reward-specific head. The FM is pretrained only using a set of state-action tuples and remains fixed during the online phase.

Our decentralized MARL for a K-device RA network is summarized as Algorithm 1 in Appendix D. Each device obeys the LBT mechanism and constantly monitors the channel. Once the channel is assessed to be clear, the device makes a transmission decision $a_{k,t}$ using its local policy conditioned on the state-action history **(Lines 6 - 11)**. Depending on the transmission result, each device updates its status. The above step is repeated for the span of T time slots, and we consider our MDP to only progress over time slots where an action is taken by the devices. Each time the devices are prepared for updating their model weights, the consensus process **(Lines 19 - 22)** first takes place for information exchange. Each consensus step consists of G rounds of communication, where weight-based averaging is performed in each round. Then, each device updates its actor and critic parameters after computing the temporal difference (TD) error **(Lines 24 - 29)**. While the actor takes the input directly from the state-action history, the critic first processes it via FM-based reward-agnostic backbone and then uses its output to update the weights of reward-specific head. As a final step, both the current state and action are stored in the history buffer.

4) Theoretical Convergence Analysis: Denoting the MDP step index by n, we establish the theoretical convergence of our decentralized multi-agent actor-critic algorithm. The required assumptions are provided in Appendix E.1.

Theorem 1 (Finite-Time Critic Convergence Rate)**.** *Consider iterative updates of $w_{k,n}$ for all $k \in \mathcal{K}$ in Algorithm 1. For any given policy π_θ and $k \in \mathcal{K}$, we have*

$$\mathbb{E}\Big[\|w_{k,n} - w_{k,n}^*\|^2\Big] \leq 2\left(\frac{c_1}{c_2}\right)^2 e^{-2c_3 G + 2c_4 n} + 2c_5(1-c_6\beta)^{n-\tau(\beta)} + 2c_7\tau(\beta)\beta, \tag{13}$$

where $c_1 = 2Kr_{max}\beta(1+\nu^{-(K-1)})$, $c_2 = (1+\gamma)\beta$, $c_3 = \ln(1-\nu^{K-1})^{-1}$, $c_4 = \ln(1+(1+\gamma)\beta)$, $\{c_5, c_6, c_7\}$ are constants independent of the step size β, and $\tau(\beta) = \mathcal{O}(\log(\beta^{-1}))$ is the mixing time.

The proof of Theorem 1 is provided in Appendix E.2. The above result establishes convergence to the TD fixed points for all devices even when rewards are shared rather than critic parameters as in

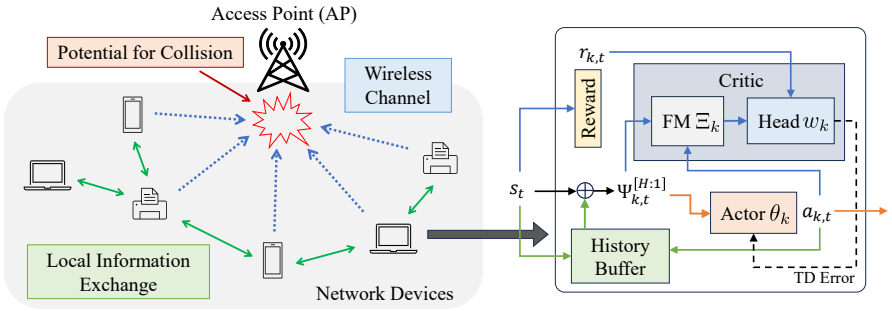


Figure 2: A visual representation of our fully decentralized MARL framework. Both the actor and critic are trained in a decentralized manner. There exist local communication links among devices that allow local information sharing for consensus.

related works [23, 24, 25]. From the first term in (13), we see that the communication rounds G for reward sharing must be sufficiently large.

Theorem 2 (Finite-Time Convergence Rate of Decentralized MARL with Local Reward Consensus)**.** *Consider the actor-critic algorithm in Algorithm 1. With step-size set as* $\alpha = \frac{1}{4L_J}$,

$$\mathbb{E}\Big[\|\nabla_\theta J(\theta^{(\hat{N})})\|^2\Big] \leq \frac{16L_J r_{max}}{N(1-\gamma)} + 18(1+\gamma)^2 \frac{\sum_{n=1}^{N}\sum_{k\in\mathcal{K}}\|w_{k,n}-w_{k,n}^*\|^2}{N}$$
$$+ 72K^3 r_{max}^2\Big((1+\nu^{-(K-1)})(1-\nu^{K-1})^G\Big)^2 + 72K(r_{max}+(1+\gamma)R_w)^2 + 18(1+\gamma)^2\xi_{approx}^{critic}, \tag{14}$$

where $\hat{N}$ *is sampled uniformly from* $\{1,\cdots,N\}$ *and* R_w *is a constant that is independent of* N.

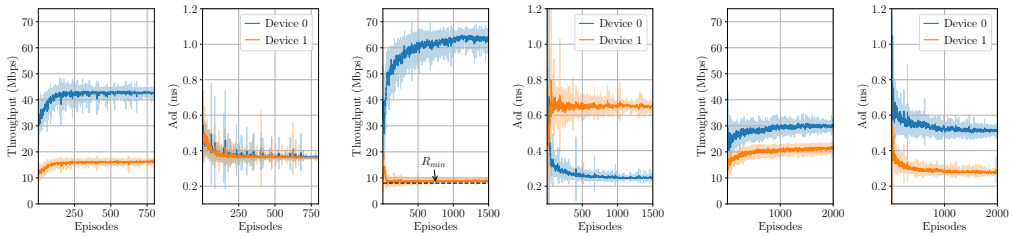
The proof of Theorem 2 is provided in Appendix E.3. Based on Theorem 2, we ensure that the output policy of Algorithm 1 converges to the neighborhood of some stationary point at a rate of $\mathcal{O}(1/N)$.

4 Numerical Evaluation

We conduct numerical experiments to evaluate our FM-aided decentralized MARL algorithm for optimizing RA-based MAC layer. The details of experimental setup are provided in Appendix C.1.

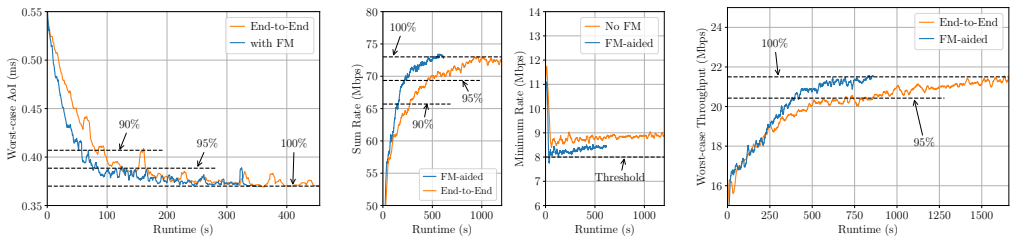
As the baselines, we first analyze the end-to-end learning behavior of our consensus-based decentralized MARL and validate MDP parameters we formulate. For ease of illustration, we show the throughput and AoI results across learning episodes with $K = 2$ devices in Fig. 3. The results in Fig. 3 confirm strong learning performance in all tasks. As expected, the fair-AoI task (Fig. 3a) successfully balances the AoI across devices. The max-rate task (Fig. 3b) effectively maximizes one device's throughput while ensuring the other remains close to $R_{\min}$. The fair-rate task (Fig. 3c) is able to jointly improve throughput of both devices while preserving the fairness. We refer readers to Appendix C.2.1 for further experiments.

With the end-to-end MARL results as proof of concepts and baselines, we are now ready to assess the effectiveness of our FM-aided MARL algorithm by comparing it with the end-to-end training. Fig. 4 presents the AoI and throughput over runtime for each downstream task. For all tasks, FM-aided learning shows superior learning efficiency, approximately improving the convergence speed by at least 30%. Due to space limitation, we relegate additional results in Appendix C.2.2. Lastly, we provide a summary of performance comparison in Table 1. We can observe that all algorithms converge to nearly the same rate and AoI. However, the proposed FM-aided MARL approach arrives at its 95%-optimal performance level *much quicker* than the end-to-end training approach. Due to space limitation, we provide our results with $K = 4$ in Appendix C.2.3, showing 55% improvement in learning speed.



(a) Fair-AoI task. (b) Max-rate task. (c) Fair-rate task.

Figure 3: Throughput and AoI performance under end-to-end learning across different RA optimization tasks, evaluated with $K = 2$ for ease of illustration.



(a) Fair-AoI task. (b) Max-rate task. (c) Fair-rate task.

Figure 4: Performance comparison of end-to-end and FM-aided MARL across different RA optimization tasks. The FM-aided approach achieves faster convergence than the end-to-end baseline.

Table 1: Performance Comparison of End-to-End and FM-aided MARL. AoI is measured for the fair-AoI task, and rate is measured for both max-rate and fair-rate tasks.

Task	Algorithm	Mean	Min.	Max.	N-Gap	95%-Time (s)
Fair-AoI	End-to-End	0.364 ± 0.007	0.363	0.364	0.002	120
(ms)	FM-aided	0.364 ± 0.002	0.362	0.366	0.013	82
Max-rate	End-to-End	36.113 ± 0.051	8.878	63.347	0.860	450
(Mbps)	FM-aided	36.443 ± 0.102	8.420	64.467	0.723	220
Fair-rate	End-to-End	25.741 ± 0.026	21.442	30.039	0.286	420
(Mbps)	FM-aided	25.659 ± 0.022	21.541	29.777	0.277	701

5 Conclusion

In this work, we proposed an FM-aided decentralized MARL framework for RA network optimization. By employing an FM as a reward-agnostic critic backbone, our algorithm effectively leverages self-supervised knowledge on network dynamics and improves learning efficiency. Compared with end-to-end training, the FM-aided approach achieves faster convergence, demonstrating its effectiveness in optimizing throughput and AoI across diverse objectives. Future work will extend this framework to more complex RA scenarios with larger action spaces (e.g., considering transmit power control).

Acknowledgments and Disclosure of Funding

This work was supported in part by the National Science Foundation (NSF) under Grant CNS2110259, Grant CNS2112471, and Grant IIS2324052; in part by the Office of Naval Research (ONR) under Grant N000142412729; and in part by the Defense Advanced Research Projects Agency (DARPA) under Grant D24AP00265 and Grant HR00112520019.

References

[1] B. Chen, J. Chen, Y. Gao, and J. Zhang, "Coexistence of LTE-LAA and Wi-Fi on 5 GHz with corresponding deployment scenarios: A survey," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 1, pp. 7–32, 2016.

[2] O. Aouedi, T.-H. Vu, A. Sacco, D. C. Nguyen, K. Piamrat, G. Marchetto, and Q.-V. Pham, "A survey on intelligent Internet of Things: Applications, security, privacy, and future directions," *IEEE Communications Surveys & Tutorials*, vol. 27, no. 2, pp. 1238–1292, 2025.

[3] N. H. Mahmood, S. Böcker, I. Moerman, O. A. López, A. Munari, K. Mikhaylov, F. Clazzer, H. Bartz, O.-S. Park, E. Mercier *et al.*, "Machine type communications: Key drivers and enablers towards the 6G era," *EURASIP Journal on Wireless Communications and Networking*, vol. 2021, no. 1, p. 134, 2021.

[4] J. C. Rodriguez, F. Grijalva, M. García, D. E. Chérrez Barragán, B. A. Acuña Acurio, and H. Carvajal, "Wireless communication technologies for smart grid distribution networks," *Engineering Proceedings*, vol. 47, no. 1, p. 7, 2023.

[5] I. Ahmad, S. Shahabuddin, H. Malik, E. Harjula, T. Leppänen, L. Loven, A. Anttonen, A. H. Sodhro, M. M. Alam, M. Juntti *et al.*, "Machine learning meets communication networks: Current trends and future challenges," *IEEE Access*, vol. 8, pp. 223 418–223 460, 2020.

[6] M. Kulin, T. Kazaz, E. De Poorter, and I. Moerman, "A survey on machine learning-based performance improvement of wireless networks: PHY, MAC and network layer," *Electronics*, vol. 10, no. 3, p. 318, 2021.

[7] X. Cao, B. Yang, C. Huang, C. Yuen, M. Di Renzo, Z. Han, D. Niyato, H. V. Poor, and L. Hanzo, "AI-assisted MAC for reconfigurable intelligent-surface-aided wireless networks: Challenges and opportunities," *IEEE Communications Magazine*, vol. 59, no. 6, pp. 21–27, 2021.

[8] M. Han, S. Khairy, L. X. Cai, Y. Cheng, and R. Zhang, "Reinforcement learning for efficient and fair coexistence between LTE-LAA and Wi-Fi," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 8, pp. 8764–8776, 2020.

[9] C. Lee, S. Park, and T. Cheong, "Dynamic-persistent CSMA: A reinforcement learning approach for multi-user channel access," *IEEE Access*, vol. 12, pp. 178 705–178 716, 2024.

[10] Y. Yu, T. Wang, and S. C. Liew, "Deep-reinforcement learning multiple access for heterogeneous wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1277–1290, 2019.

[11] Y. Yu, S. C. Liew, and T. Wang, "Multi-agent deep reinforcement learning multiple access for heterogeneous wireless networks with imperfect channels," *IEEE Transactions on Mobile Computing*, vol. 21, no. 10, pp. 3718–3730, 2022.

[12] L. Zhang, H. Yin, Z. Zhou, S. Roy, and Y. Sun, "Enhancing WiFi multiple access performance with federated deep reinforcement learning," in *Proceedings of IEEE Vehicular Technology Conference*. IEEE, 2020, pp. 1–6.

[13] Z. Guo, Z. Chen, P. Liu, J. Luo, X. Yang, and X. Sun, "Multi-agent reinforcement learning-based distributed channel access for next generation wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 5, pp. 1587–1599, 2022.

[14] Y. He, X. Gang, and Y. Gao, "Intelligent decentralized multiple access via multi-agent deep reinforcement learning," in *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2024, pp. 1–6.

[15] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," *arXiv preprint arXiv:2312.00752*, 2023.

[16] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized gossip algorithms," *IEEE Transactions on Information Theory*, vol. 52, no. 6, pp. 2508–2530, 2006.

[17] R. S. Sutton, A. G. Barto *et al.*, *Reinforcement learning: An introduction*. MIT Press Cambridge, 1998.

[18] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proceedings of International Conference on Machine Learning (ICML)*. PMLR, 2018, pp. 1861–1870.

[19] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[20] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[21] Y. Sun, S. Ma, R. Madaan, R. Bonatti, F. Huang, and A. Kapoor, "SMART: Self-supervised multi-task pretraining with control transformers," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2023.

[22] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. New York, NY, USA: Cambridge University Press, 2005.

[23] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar, "Fully decentralized multi-agent reinforcement learning with networked agents," in *Proceedings of International Conference on Machine Learning (ICML)*. PMLR, 2018, pp. 5872–5881.

[24] Z. Chen, Y. Zhou, R.-R. Chen, and S. Zou, "Sample and communication-efficient decentralized actor-critic algorithms with finite-time analysis," in *Proceedings of International Conference on Machine Learning (ICML)*. PMLR, 2022, pp. 3794–3834.

[25] F. Hairi, J. Liu, and S. Lu, "Finite-time convergence and sample complexity of multi-agent actor-critic reinforcement learning with average reward," in *Proceedings of International Conference on Learning Representations (ICLR)*. PMLR, 2022.

[26] J. Du, T. Lin, C. Jiang, Q. Yang, C. F. Bader, and Z. Han, "Distributed foundation models for multi-modal learning in 6G wireless networks," *IEEE Wireless Communications*, vol. 31, no. 3, pp. 20–30, 2024.

[27] J. Fontaine, A. Shahid, and E. De Poorter, "Towards a wireless physical-layer foundation model: Challenges and strategies," in *2024 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2024, pp. 1–7.

[28] Z. Chen, Z. Zhang, and Z. Yang, "Big AI models for 6G wireless networks: Opportunities, challenges, and research directions," *IEEE wireless communications*, vol. 31, no. 5, pp. 164–172, 2024.

[29] Z. Yang, H. Du, D. Niyato, X. Wang, Y. Zhou, L. Feng, F. Zhou, W. Li, and X. Qiu, "Revolutionizing wireless networks with self-supervised learning: A pathway to intelligent communications," *IEEE Wireless Communications*, 2025.

[30] T. Yang, P. Zhang, M. Zheng, Y. Shi, L. Jing, J. Huang, and N. Li, "WirelessGPT: A generative pre-trained multi-task learning framework for wireless communication," *IEEE Network*, 2025.

[31] Y. Sheng, J. Wang, X. Zhou, L. Liang, H. Ye, S. Jin, and G. Y. Li, "A wireless foundation model for multi-task prediction," *arXiv preprint arXiv:2507.05938*, 2025.

[32] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[33] R. Srikant and L. Ying, "Finite-time error bounds for linear stochastic approximation and td learning," in *Proceedings of Conference on Learning Theory*. PMLR, 2019, pp. 2803–2830.

[34] T. Xu, Z. Wang, and Y. Liang, "Improving sample complexity bounds for (natural) actor-critic algorithms," in *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020, pp. 4358–4369.

A Related Work

A.1 MARL-based Random Access (RA) Network Optimization

To address the challenges present in RA network optimization, several reinforcement learning (RL)-based approaches have been proposed. For example, authors in [8] proposed Q-learning-based contention window selection algorithms for each cooperative and non-cooperative setting to maximize the total throughput while satisfying the fairness constraints. In [9], deep RL based on SAC and long short-term memory (LSTM) models was utilized to dynamically adjust the device waiting time and optimize the network throughput. Another approach in RA optimization is to develop a deterministic transmission policy for each participating device, for which several MARL-based strategies have

been proposed. In [10], a deep Q-network was adopted to make transmission decisions for each RA device with an aim to maximize the generalized α-fairness objective. This approach was later extended to account for an imperfect wireless channel in which feedback signals for information collection can be corrupted [11]. The work in [12] employed a federated learning framework to implement distributed policy learning in RA networks, where each device is equipped with a DNN for decision-making. Furthermore, QMIX and multi-agent PPO algorithms were explored in [13] and [14], respectively, to implement MARL-based RA and improve network performance.

A.2 Foundation Model (FM) and its Application in Wireless Communications

FMs have revolutionized a wide range of AI-driven domains including natural language processing and computer vision. By leveraging large-scale data and powerful representation learning, FM-enhanced tasks in these areas have demonstrated remarkable advances in reasoning, perception, and generative capabilities.

Building on these successes, FMs are expected to play a significant role in data-driven optimization strategies for wireless communications, spanning diverse PHY-layer and MAC-layer tasks [26, 27, 28, 29]. When combined with emerging communication paradigms, such as semantic communications and integrated sensing and communications (ISAC), FMs can unlock immense potential. In particular, their ability to generalize across tasks and domains makes them well-suited for addressing the challenges of dynamic wireless environments, enabling robust adaptability, cross-layer optimization, and real-time decision-making. Moreover, FMs can drive a paradigm shift in designing scalable, efficient, and intelligent wireless networks by effectively unifying domain-specific communication frameworks.

There exist several early works on developing FMs specified for wireless communications and networks. For example, a FM called WirelessGPT is introduced in [30], which is tailored for wireless communication and sensing tasks. In [31], a patch-masked FM is developed for multi-task prediction in wireless networks, demonstrating effectiveness in channel, angle, traffic prediction tasks.

B Actor-Critic Framework for MARL

In the actor-critic method, MARL is implemented using two sets of neural networks: *actors* and *critics*. Each device k is equipped with an actor network parameterized by weights θ_k, which defines its local policy π_{θ_k}. Let $\theta = [\theta_1^\top, \theta_2^\top, \ldots, \theta_K^\top]^\top$ denote the joint weight vector of all K actors. Denoting the MDP step index by n, the goal of MARL is to optimize θ to maximize the state-value function under the joint policy π_θ, which is defined as $V_\theta(s) = \mathbb{E}_{\pi_\theta}\left[\sum_{n=0}^{\infty} \gamma^n \overline{r}_n \middle| s_0 = s\right]$, where $\overline{r}_n = \frac{1}{K}\sum_{k=1}^{K} r_{k,n}$ is the global average reward, and $\gamma \in [0,1]$ is the discount factor. For each device k, the state-value function is typically estimated using a critic network of parameters w_k. Applying the Bellman equation [17], the function can be expressed as $V_{w_k}(s) = \mathbb{E}_{\pi_\theta}\left[\overline{r} + \gamma V_{w_k}(s')\right]$. During training, each actor selects an action based on its current policy, and the corresponding critic evaluates the resulting reward. The actor policy is then updated using the policy gradient $\mathbb{E}_{s,a}[\nabla_{\theta_k} \log \pi_{\theta_k}(a_k|s) \cdot \text{Adv}_\theta(s, \boldsymbol{a})]$, where $\text{Adv}_\theta(s, \boldsymbol{a}) = \bar{r} + \gamma V_\theta(s') - V_\theta(s)$ is the advantage function. This update encourages the actor to select actions that yield higher expected future rewards relative to the baseline value.

C Numerical Evaluation

C.1 Experimental Settings

We consider $K = \{2, 4\}$ devices participating in RA over $T = 600$ time slots. As described in Sect. 2, all devices operate under the LBT mechanism. Following the IEEE 802.11 protocol, the SIFS and DIFS durations are set to 2 and 4 time slots, respectively, with each slot lasting $T_{\mathsf{s}} = 9\ \mu s$. Each data packet transmission requires $\tau = 10$ time slots, while ACK signals occupy 4 time slots.

We set the number of AP antennas to $N_{\mathsf{r}} = 2$, the bandwidth to $B = 10$ MHz, and the transmit power to $p_k = 0$ dBm, with the noise spectral density fixed at -174 dBm/Hz. For a given number of devices K, device-to-AP distances are uniformly distributed between 2 m and 14 m. For example, with

$K = 4$, the device distances are $\{d_{k,t}\}_{k=1}^{K=4} = \{2, 6, 10, 14\}$. The distance-dependent large-scale fading factor $\eta_{k,t}$ is modeled as $\eta_{k,t} = L_{\mathsf{o}} \left(\frac{d_{k,t}}{d_{\mathsf{o}}}\right)^{z}$, where $L_{\mathsf{o}} = -40$ dB is the reference pathloss, $d_{\mathsf{o}} = 1$ m is the reference distance, and $z = 4$ is the pathloss exponent.

The network graph $\mathcal{G}$ is generated using the Watts-Strogatz model [32], where each device connects to one neighboring device with zero rewiring probability. To generate the consensus weight matrix $\mathbf{C}$, we assign equal weights are assigned to each device's established links, i.e., for each device $k \in \mathcal{K}$, $c_{kk'} = \frac{1}{|\mathcal{N}_k|+1}, \forall k' \in \mathcal{N}_k$, where $\mathcal{N}_k$ denotes the neighbor set of device k. Each consensus step consists of $G = 3$ communication rounds.

For the critic-backbone FM, we set the embedding size to $N_{\mathsf{e}} = 64$ such that the encoders, transformer blocks, and prediction head all generate embeddings of dimension 64. Both the state and action encoders are implemented as two-layer MLPs with width 64. The FM consists of four transformer blocks, each containing a self-attention layer with 8 heads and an MLP of width $4N_{\mathsf{e}}$ and depth 1. The prediction head is a linear transformation of size $2N_{\mathsf{e}} \times N_{\mathsf{e}}$ with bias. No dropout is applied in the FM design.

Each device's FM is pretrained using more than $100,000$ reward-independent RA network samples, each consisting of a tuple $\{t, s_t, a_{k,t}\}$. We employ stochastic gradient descent (SGD) with a learning rate of 0.005, applying 500 updates with a batch size 256. The target encoder is updated using a soft update with rate 0.005. As an illustrative example, we present the FM pretraining loss curve for the case of $K = 4$ in Fig. 5.

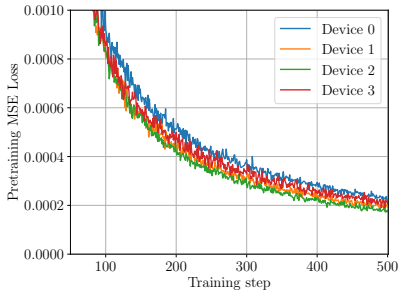


Figure 5: FM pretraining loss curve with $K = 4$.

For both actor and reward-specific critic head, we adopt a ResNet block of width 128 and depth 3. The actor and critic weights are updated using the SGD optimizer, with learning rates $\alpha = 0.002$ and $\beta = 0.001$ for $K = 2$, and $\alpha = 0.001$ and $\beta = 0.0005$ for $K = 4$. Discount factor is set to $\gamma = 0.3$. For the end-to-end training baseline, we employ a ResNet block of width 128 and depth 7 for both actor and critic to ensure a comparable number of parameters for fair evaluation. Since the state dimension grows proportionally with K, we set $H = 10$ for $K = 2$ and $H = 6$ for $K = 4$ to stabilize the input dimension of the actor–critic network.

For the MDP parameters, we set $\lambda = 60$, $\omega_{\mathsf{FA}} = 15$, $\omega_{\mathsf{MR}} = 24 \times 10^6$, $\omega_{\mathsf{P1}} = 0.5$, $\omega_{\mathsf{P2}} = 8$, $\omega_{\mathsf{FR1}} = 256 \times 10^6$, $\omega_{\mathsf{FR2}} = 0.7$, and $\alpha = 12$ for parameter scaling. These values are heuristically selected to bound the reward to a reasonable range. All results are averaged over at least 20 independent runs.

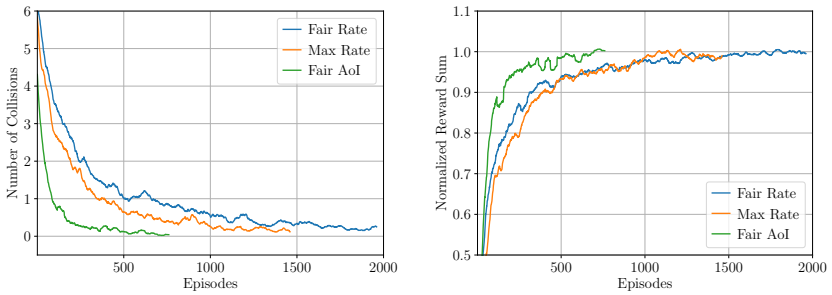
As performance metrics on RA-based networks, we evaluate throughput (in Mbps) and AoI (in ms). In addition, we assess fairness across devices using the normalized gap (N-Gap) defined as

$$\text{N-Gap of } \{x\} = \frac{\max(\{x\}) - \min(\{x\})}{\max(\{x\})}. \tag{15}$$

C.2 Additional Experimental Results

C.2.1 Performance Analysis on Downstream Tasks in Terms of Collision Frequency and Reward

In Fig. 6, we present two additional performance metrics for evaluating our consensus-based decentralized MARL in an end-to-end learning setting across different RA optimization tasks: the number of collisions (Fig. 6a) and the normalized reward sum (Fig. 6b). For readability, all results are smoothed using a moving average over 40 samples. The results show that the fair-AoI task reduces collisions the fastest, suggesting that balancing AoI is the simplest among the three tasks. This is intuitive, as the fair-AoI objective does not require consideration of channel conditions, which directly influence device data rates. By contrast, the fair-rate task exhibits the highest collision counts even after many episodes. This is because its reward function imposes stronger penalties for rate unfairness across devices. Consequently, the policy that reduces collisions emerges later in training, consistent with Fig. 3c where device rates increase gradually while maintaining fairness gaps.

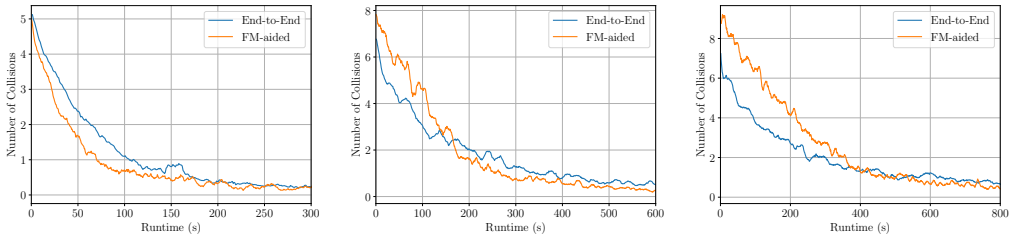


(a) Number of Collisions. (b) Normalized reward sum.

Figure 6: Collision and reward performance under end-to-end learning across different RA optimization tasks, evaluated with $K = 2$ for ease of illustration.

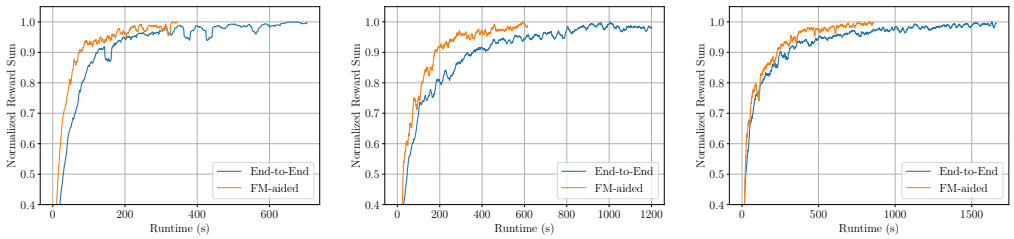
C.2.2 Comparison of FM-aided and End-to-End Learning in Collision Frequency and Reward Evaluation

In Figs. 7 and 8, we compare the collision frequency and reward sum between end-to-end and FM-aided MARL across different RA optimization tasks. Consistent with the trend observed in Fig. 4, our algorithm demonstrates superior performance, achieving faster collision reduction and quicker reward maximization. The performance gap in collision frequency is particularly evident in the fair-AoI and max-rate tasks, whereas it appears less pronounced in the fair-rate task. This is consistent with our analysis in Appendix C.2.1, where the learning is shown to focus on rate fairness first. Regarding reward sum, FM-aided MARL converges in roughly half the time required by end-to-end learning. Overall, FM-aided MARL consistently outperforms end-to-end learning in terms of learning speed.



(a) Fair AoI task. (b) Max rate task. (c) Fair rate task.

Figure 7: Collision frequency comparison between end-to-end and FM-aided MARL.

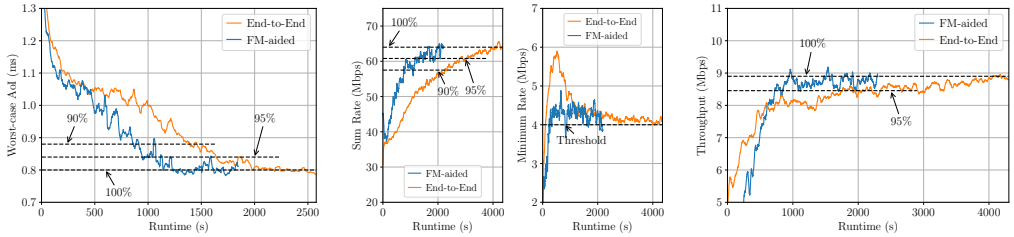


(a) Fair AoI task. (b) Max rate task. (c) Fair rate task.

Figure 8: Normalized reward sum comparison between end-to-end and FM-aided MARL.

C.2.3 Performance Comparison of End-to-End and FM-aided MARL across Different RA Optimization Tasks with 4 Devices

We evaluate the effectiveness of our FM-aided MARL algorithm by comparing it with conventional end-to-end training for $K = 4$. We present the evolution of AoI and throughput over runtime for each downstream task in Fig. 9, and the corresponding results are summarized in Table 2. As the number of devices increases, overall throughput decreases and AoI rises due to heightened competition for channel access. Nevertheless, the observed trends remain consistent with the $K = 2$ case, where FM-aided MARL substantially increases learning speed. In particular, the runtime required to achieve 95%-performance is reduced by approximately 55%, which is a much increased improvement than 30% for the $K = 2$ case. These results suggest that leveraging self-supervised FMs becomes increasingly beneficial in more complex learning environments.



(a) Fair AoI task. (b) Max rate task. (c) Fair rate task.

Figure 9: Performance comparison of end-to-end and FM-aided MARL across three RA optimization tasks with $K = 4$. The FM-aided MARL achieves faster convergence than the end-to-end baseline.

Table 2: Performance Comparison of End-to-End and FM-aided MARL with $K = 4$. AoI is measured for the fair-AoI task, and rate is measured for both max-rate and fair-rate tasks.

Task	Algorithm	Mean	Min.	Max.	N-Gap	95%-Time (s)
Fair AoI	End-to-End	0.742 ± 0.014	0.732	0.754	0.028	1645
(ms)	FM-aided	0.757 ± 0.025	0.745	0.773	0.037	760
Max Rate	End-to-End	16.026 ± 0.047	4.567	49.637	0.793	3000
(Mbps)	FM-aided	15.861 ± 0.113	4.332	45.937	0.739	1300
Fair Rate	End-to-End	10.600 ± 0.026	8.888	13.083	0.321	2340
(Mbps)	FM-aided	11.530 ± 0.025	8.791	14.188	0.277	1055

D Algorithm Pseudocode

Below is the pseudocode for our FM-aided consensus-based decentralized MARL algorithm.

Algorithm 1: FM-aided Decentralized MARL for RA Network Optimization.

Input: device set $\mathcal{K}$, neighbor sets $\{\mathcal{N}_k\}_{k\in\mathcal{K}}$, pretrained FM weights $\{\Xi_k\}_{k\in\mathcal{K}}$, time horizon length T, history length H, consensus weight matrix $\mathbf{C}$, consensus iteration count G, actor rate α, critic rate β,

Initialize: actor weights θ_k, critic weights w_k, transmit flag $\mathrm{f}_k = \text{False}$, and ready-for-update status $u_k = \text{False}$ for all $k \in \mathcal{K}$

Load: FM weights Ξ_k for all $k \in \mathcal{K}$

for $t = 1, 2 \ldots, T$ **do**

  for $k \in \mathcal{K}$ **do**

    Update q_t, f_k, and $\ell_{k,t}$

    if $q_t = 0$ **then**

      Acquire state s_t and reward $r_{k,t}$

      Select $a_{k,t} \sim \pi_{\theta_k}(\cdot\,|\Psi_{k,t}^{[H:1]})$

      if $a_{k,t} = 1$ **then**

        $\mathrm{f}_k \leftarrow \text{True}$

      $u_k \leftarrow \text{True}$

  for $k \in \mathcal{K}$ **do**

    if $f_k = \textit{True}$ **then**

      $q_t \leftarrow 1$

      Transmit a packet

    if *ACK received* **then**

      $\ell_{k,t} \leftarrow 0$

  if $u_k = \textit{True}, \forall k \in \mathcal{K}$ **then**

    $\tilde{r}_{k,0} \leftarrow r_{k,t}$ for all $k \in \mathcal{K}$

    for $g = 1, 2 \ldots, G$ **do**

      $\tilde{r}_{k,g} \leftarrow \sum_{k'\in\mathcal{N}_k} c_{kk'}\tilde{r}_{k',g-1}$ for all $k \in \mathcal{K}$

    $\tilde{r}_{k,t} \leftarrow \tilde{r}_{k,G}$ for all $k \in \mathcal{K}$

    for $k \in \mathcal{K}$ **do**

      $\delta_k \leftarrow \tilde{r}_{k,t} + \gamma V_{\Xi_k w_k}(\Psi_{k,t}^{[H:1]}) - V_{\Xi_k w_k}(\Psi_{k,\tilde{t}_k^{[1]}}^{[H+1:2]})$

      $w_k \leftarrow w_k - \beta\delta_k \cdot \nabla V_{\Xi_k w_k}(\Psi_{k,\tilde{t}_k^{[1]}}^{[H+1:2]})$

      $\delta_k \leftarrow \tilde{r}_{k,t} + \gamma V_{\Xi_k w_k}(\Psi_{k,t}^{[H:1]}) - V_{\Xi_k w_k}(\Psi_{k,\tilde{t}_k^{[1]}}^{[H+1:2]})$

      $\theta_k \leftarrow \theta_k + \alpha\delta_k \cdot \nabla \log \pi_{\theta_k}(a_{k,\tilde{t}_k^{[1]}}|\Psi_{k,\tilde{t}_k^{[1]}}^{[H+1:2]})$

      Store s_t and $a_{k,t}$ in the history buffer

      $u_k \leftarrow \text{False}$

Output: θ_k for all $k \in \mathcal{K}$

E Theoretical Convergence Analysis

E.1 Assumptions

Assumption 1 (Bounded Reward)**.** *For each downstream task, there exists a positive constant r_{max} such that $r_{k,n} \in [-r_{max}, r_{max}]$ for any $n \geq 0$ and $k \in \mathcal{K}$.*

Assumption 2 (Mixing Time)**.** *There exist a stationary distribution ζ for (s, a), and positive constants κ and $\rho \in (0, 1)$, such that $\sup_{s\in\mathcal{S}} \|P(s_n, a_n|s_0 = s) - \zeta(\theta)\|_{TV} \leq \kappa\rho^n, \forall n \geq 0$.*

Assumption 3 (Lipschitz Continuity)**.** *$J(\theta)$ is L_J-Lipschitz continuous w.r.t. θ, i.e., there exists a positive constant L_J such that, for any θ and θ', we have $|J(\theta) - J(\theta')| \leq L_J\|\theta - \theta'\|_2$.*

Assumption 4 (Consensus Matrix)**.** *The consensus weight matrix $\mathbf{C}$ is doubly stochastic. Additionally, for all $k, k' \in \mathcal{K}$, there exists a positive constant $\nu > 0$ such that (i) $c_{kk} \geq \nu$ and (ii) $c_{kk'} \geq \nu$ whenever devices k and k' are connected.*

E.2 Proof of Theorem 1

For the convenience of notation, we use $\Psi_{k,n}$ and $\Psi_{k,n+1}$ to represent $\Psi_{k,\tilde{t}_k^{[1]}}^{[H+1:2]}$ and $\Psi_{k,t}^{[H:1]}$ in Algorithm 1, respectively. We first make the following assumption as in [23, 25].

Assumption 5. *State-value function of each $k \in \mathcal{K}$ can be approximated using a linear function, i.e., $V_{\Xi_k w_k}(\Psi_{k,n}) = \phi_{\Xi_k}(\Psi_{k,n})^\top w_k$ where $\phi_{\Xi_k}(\Psi_{k,n})$ is the uniformly bounded feature associated with $\Psi_{k,n}$, i.e., $\|\phi_{\Xi_k}(\Psi_{k,n})\| \leq 1$.*

Let us define $w_{k,n}^*$ to be the optimal weights for device k's critic at a step index n. To derive the upper bound on the difference between $w_{k,n}$ and $w_{k,n}^*$, i.e., $\|w_{k,n} - w_{k,n}^*\|$, we separate the difference into $w_{k,n} - \bar{w}_{k,n}$ and $\bar{w}_{k,n} - w_{k,n}^*$, where $\bar{w}_{k,n} \triangleq \frac{1}{K}\sum_{k\in\mathcal{K}} w_{k,n}$ and derive the bound on each of them.

We start on the first part of our theorem. Recall that each gradient update step yields

$$w_{k,n} = w_{k,n-1} + \beta\left(\tilde{r}_{k,n-1} + \gamma\phi^\top(\Psi_{k,n})w_{k,n-1} - \phi^\top(\Psi_{k,n-1})w_{k,n-1}\right)\phi(\Psi_{k,n-1}) \tag{16}$$

$$\bar{w}_{k,n} = \bar{w}_{k,n-1} + \beta\left(\bar{r}_{n-1} + \gamma\phi^\top(\Psi_{k,n})\bar{w}_{k,n-1} - \phi^\top(\Psi_{k,n-1})\bar{w}_{k,n-1}\right)\phi(\Psi_{k,n-1}) \tag{17}$$

where $\tilde{r}_{k,n} = [\mathbf{C}^G]_k r_n$, $\bar{r}_n = \frac{1}{K}\mathbf{1}^\top r_n$, and $r_n = [r_{1,n}, \cdots, r_{K,n}]^\top$. Note that $[\mathbf{C}^G]_k$ denotes the k-th row of matrix $\mathbf{C}^G$. We get the consensus error vector given by

$$\begin{aligned}
e_{k,n} &= w_{k,n} - \bar{w}_{k,n} && (18)\\
&= (w_{k,n-1} - \bar{w}_{k,n-1}) + \beta\phi(\Psi_{k,n-1})\left(\left[[\mathbf{C}^G]_k - \frac{1}{K}\mathbf{1}^\top\right]r_{n-1}\right)\\
&\qquad + \beta\phi(\Psi_{k,n-1})[\gamma\phi(\Psi_{k,n}) - \phi(\Psi_{k,n-1})]^\top(w_{k,n-1} - \bar{w}_{k,n-1}) && (19)\\
&= (w_{k,n-1} - \bar{w}_{k,n-1}) + \beta\phi(\Psi_{k,n-1})r_{n-1}^\top\left[[\mathbf{C}^G]_k^\top - \frac{1}{K}\mathbf{1}\right]\\
&\qquad + \beta\phi(\Psi_{k,n-1})[\gamma\phi(\Psi_{k,n}) - \phi(\Psi_{k,n-1})]^\top(w_{k,n-1} - \bar{w}_{k,n-1}) && (20)\\
&= e_{k,n-1} + \beta C_{k,n-1}\left[[\mathbf{C}^G]_k^\top - \frac{1}{K}\mathbf{1}\right] + \beta B_{k,n-1}e_{k,n-1} && (21)\\
&= (\mathbf{I} + \beta B_{k,n-1})e_{k,n-1} + \beta C_{k,n-1}\left[[\mathbf{C}^G]_k^\top - \frac{1}{K}\mathbf{1}\right], && (22)
\end{aligned}$$

where $B_{k,n} = \phi(\Psi_{k,n})[\gamma\phi(\Psi_{k,n+1}) - \phi(\Psi_{k,n})]^\top$ and $C_{k,n} = \phi(\Psi_{k,n})r_n^\top$. Note that (22) is a function of $e_{k,n-1}$. Hence, we can express $e_{k,n}$ in an iterative form:

$$e_{k,n} = \left[\prod_{x=0}^{n-1}(\mathbf{I} + \beta B_{k,x})\right]e_{k,0} + \beta\sum_{x=0}^{n-1}\left[\prod_{y>x}^{n-1}(\mathbf{I} + \beta B_{k,y})\right]C_{k,x}\left[[\mathbf{C}^G]_k^\top - \frac{1}{K}\mathbf{1}\right]. \tag{23}$$

Since $e_{k,0}$ is zero due to $w_{k,0} = \bar{w}_{k,0}$, we can express the norm of $e_{k,n}$ as

$$\begin{aligned}
\|e_{k,n}\| &= \left\|\beta\sum_{x=0}^{n-1}\Big[\prod_{y>x}^{n-1}(\mathbf{I} + \beta B_{k,y})\Big]C_{k,x}\left[[\mathbf{C}^G]_k^\top - \frac{1}{K}\mathbf{1}\right]\right\| && (24)\\
&\leq \beta\sum_{x=0}^{n-1}\left\|\prod_{y>x}^{n-1}(\mathbf{I} + \beta B_{k,y})\right\| \cdot \|C_{k,x}\| \cdot \left\|[\mathbf{C}^G]_k^\top - \frac{1}{K}\mathbf{1}\right\|. && (25)
\end{aligned}$$

We bound each term in (25) as follows. For the first term, we have

$$\begin{aligned}
\left\|\prod_{y>x}^{n-1}(\mathbf{I} + \beta B_{k,y})\right\| &\leq \prod_{y>x}^{n-1}\|\mathbf{I} + \beta B_{k,y}\| \leq \prod_{y>x}^{n-1}(\|\mathbf{I}\| + \|\beta B_{k,y}\|)\\
&\leq \prod_{y>x}^{n-1}(1 + \beta(1+\gamma)) && (26)\\
&= (1 + \beta(1+\gamma))^{n-1-x}, && (27)
\end{aligned}$$

where the last inequality is due to Assumption 5. For the second term, using Assumptions 1 and 5, we have

$$\|C_{k,n}\| = \left\|\phi(\Psi_{k,n})[r_{1,n}, \cdots, r_{K,n}]\right\| \tag{28}$$
$$\leq \|\phi(\Psi_{k,n})\| \cdot \left\|[r_{1,n}, \cdots, r_{K,n}]\right\| \tag{29}$$
$$\leq \left\|[r_{1,n}, \cdots, r_{K,n}]\right\| \tag{30}$$
$$\leq \sqrt{K} r_{\max}. \tag{31}$$

For the third term, we get

$$\left\|[\mathbf{C}^G]_k^\top - \frac{1}{K}\mathbf{1}\right\| \leq 2\sqrt{K}\frac{(1+\frac{1}{\nu^{(K-1)}})}{1-\nu^{K-1}}(1-\nu^{K-1})^{G+1} \tag{32}$$
$$= 2\sqrt{K}(1+\nu^{-(K-1)})(1-\nu^{K-1})^G. \tag{33}$$

By combining each term, we obtain

$$\beta \sum_{x=0}^{n-1} \left\| \prod_{y>x}^{n-1} (\mathbf{I} + \beta B_y) \right\| \cdot \|C_x\| \cdot \left\|[\mathbf{C}^G]_k^\top - \frac{1}{K}\mathbf{1}\right\|$$
$$\leq \beta \sum_{x=0}^{n-1} (1+\beta(1+\gamma))^{n-1-x} \cdot \sqrt{K} r_{\max} \cdot 2\sqrt{K}(1+\nu^{-(K-1)})(1-\nu^{K-1})^G \tag{34}$$
$$= 2K r_{\max} \beta (1+\nu^{-(K-1)})(1-\nu^{K-1})^G \sum_{x=0}^{n-1} (1+\beta(1+\gamma))^{n-1-x} \tag{35}$$

Thus, the upper bound on $\|w_{k,n} - \bar{w}_{k,n}\|$ becomes

$$\|w_{k,n} - \bar{w}_{k,n}\| \leq 2K r_{\max} \beta (1+\nu^{-(K-1)})(1-\nu^{K-1})^G \sum_{x=0}^{n-1} (1+\beta(1+\gamma))^{n-1-x} \tag{36}$$

To further understand (36), we simplify its right-hand-side (RHS). Let us first denote $c_1 \triangleq 2K r_{\max}\beta(1+\nu^{-(K-1)})$ and consider $(1-\nu^{K-1})^G = e^{-c_3 G}$ with $c_3 \triangleq \ln(1-\nu^{K-1})^{-1} > 0$. By denoting $c_2 \triangleq (1+\gamma)\beta > 0$, we also have

$$\sum_{x=0}^{n-1} (1+(1+\gamma)\beta)^{n-1-x} = \sum_{x=0}^{n-1} (1+c_2)^x = \frac{(1+c_2)^n - 1}{c_2} \leq \frac{(1+c_2)^n}{c_2}.$$

Furthermore, we have $(1+c_2)^n = e^{n\ln(1+c_2)} = e^{c_4 n}$ where $c_4 \triangleq \ln(1+c_2)$. Combining the points made above, (36) becomes

$$\|w_{k,n} - \bar{w}_{k,n}\| \leq \frac{c_1}{c_2} e^{-c_3 G + c_4 n}. \tag{37}$$

Note that $\frac{c_1}{c_2}$ is a constant independent of step size β. The inequality in (37) indicates that if the exponent $-c_3 G + c_4 n$ remains a sufficiently large negative number, the consensus error should be sufficiently small.

We now work on the second part of our theorem, which is on the convergence of average parameter. Using [33], we have

Lemma 1. *(Theorem 7 of [33]) For any $n > \tau(\beta)$ and for sufficiently small constant step size β, the finite-time convergence bound for average parameter is*

$$\mathbb{E}\left[\|\bar{w}_{k,n} - w^*_{k,n}\|^2\right] \leq c_5(1-c_6\beta)^{n-\tau(\beta)} + c_7\tau(\beta)\beta \tag{38}$$

where c_5, c_6, c_7 are constants independent of step size β, and $\tau(\beta) = \mathcal{O}(\log(\frac{1}{\beta}))$ is mixing time.

By Remark 1 in [33], $\beta\tau(\beta) \to 0$ as $\beta \to 0$.

We are now ready to bound $\|\bar{w}_{k,n} - w^*_{k,n}\|$. For $k \in \mathcal{K}$, we have

$$\mathbb{E}\left[\|\bar{w}_{k,n} - w^*_{k,n}\|^2\right] \leq 2\mathbb{E}\left[\|w_{k,n} - \bar{w}_{k,n}\|^2\right] + 2\mathbb{E}\left[\|\bar{w}_{k,n} - w^*_{k,n}\|^2\right] \tag{39}$$
$$\leq 2\left(\frac{c_1}{c_2}\right)^2 e^{-2c_3 G + 2c_4 n} + 2c_5(1-c_6\beta)^{n-\tau(\beta)} + 2c_7\tau(\beta)\beta. \tag{40}$$

Note that, in the finite time result above, the only constant that is dependent on β is c_4.

E.3 Proof of Theorem 2

For the ease of notation, we define $v_{k,n}(w_{k,n}) = \tilde{\delta}_{k,n} \cdot \psi_{k,n}$ and $h_{k,n}(w_{k,n}) = \delta^*_{k,n} \cdot \psi_{k,n}$, where $\psi_{k,n} = \nabla \log \pi_{\theta_k}(a_{k,n}|\Psi_{k,n})$, $\tilde{\delta}_{k,n}$ is the TD error computed using $\tilde{r}_{k,n} = [\mathbf{C}^G]_k[r_{1,n}, \cdots, r_{K,n}]^\top$, and $\delta^*_{k,n}$ is the TD error computed using $\bar{r}_t = \frac{1}{K}\mathbf{1}^\top[r_{1,n}, \cdots, r_{K,t}]^\top$. We also define $w_n = [w_{1,n}, \ldots, w_{k,n}]$, $v_n(w_n) = [v_{1,n}(w_{1,n}), \ldots, v_{k,n}(w_{k,n})]$, and $h_n(w_n) = [h_{1,n}(w_{1,n}), \ldots, h_{k,n}(w_{k,n})]$. Lastly, we define

$$\begin{aligned}\text{Adv}_{w_n}(s_n, \boldsymbol{a}_n) &= \mathbb{E}_{s'\sim P(\cdot|s,\boldsymbol{a}), r\sim d_r(s,\boldsymbol{a})}[\delta_{w_n}(s, \boldsymbol{a}, s')|s = s_n, \boldsymbol{a} = \boldsymbol{a}_n] &(41)\\ &= \mathbb{E}_{s'\sim P(\cdot|s,\boldsymbol{a}), r\sim d_r(s,\boldsymbol{a})}[r + \gamma V_{w_n}(s') - V_{w_n}(s)|s = s_n, \boldsymbol{a} = \boldsymbol{a}_n] &(42)\end{aligned}$$

and

$$g(w_n, \theta^{(n)}) = \mathbb{E}[\text{Adv}_{w_n}(s_n, \boldsymbol{a}_n)\psi_{\theta^{(n)}}(s_n, \boldsymbol{a}_n)] \tag{43}$$

We now make the following assumption on $\psi_{k,n}$.

Assumption 6. *For any policy parameter θ_k, the score function $\psi_{k,n}$ is uniformly bounded, i.e., $\|\psi_{k,n}\|^2 \leq 1$.*

Since $J(\theta)$ is L_J-Lipschitz continuous from Assumption 3, we can apply descent lemma to obtain the following result:

$$\begin{aligned}J(\theta^{(n+1)}) &\geq J(\theta^{(n)}) + \langle \nabla_\theta J(\theta^{(n)}), \theta^{(n+1)} - \theta^{(n)}\rangle - \frac{L_J}{2}\|\theta^{(n+1)} - \theta^{(n)}\|^2 &(44)\\ &= J(\theta^{(n)}) + \alpha\langle \nabla_\theta J(\theta^{(n)}), v_n(w_n) - \nabla_\theta J(\theta^{(n)}) + \nabla_\theta J(\theta^{(n)})\rangle - \frac{L_J\alpha^2}{2}\|v_n(w_n)\|^2 &(45)\\ &= J(\theta^{(n)}) + \alpha\|\nabla_\theta J(\theta^{(n)})\|^2 + \alpha\langle \nabla_\theta J(\theta^{(n)}), v_n(w_n) - \nabla_\theta J(\theta^{(n)})\rangle\\ &\qquad - \frac{L_J\alpha^2}{2}\|v_n(w_n) - \nabla_\theta J(\theta^{(n)}) + \nabla_\theta J(\theta^{(n)})\|^2 &(46)\\ &\geq J(\theta^{(n)}) + \left(\frac{1}{2}\alpha - L_J\alpha^2\right)\|\nabla_\theta J(\theta^{(n)})\|^2\\ &\qquad - \left(\frac{1}{2}\alpha + L_J\alpha^2\right)\|v_n(w_n) - \nabla_\theta J(\theta^{(n)})\|^2, &(47)\end{aligned}$$

where the last inequality is due to

$$\langle \nabla_\theta J(\theta^{(n)}), v_n(w_n) - \nabla_\theta J(\theta^{(n)})\rangle \geq -\frac{1}{2}\|\nabla_\theta J(\theta^{(n)})\|^2 - \frac{1}{2}\|v_n(w_n) - \nabla_\theta J(\theta^{(n)})\|^2 \tag{48}$$

and

$$\|v_n(w_n) - \nabla_\theta J(\theta^{(n)}) + \nabla_\theta J(\theta^{(n)})\|^2 \leq 2\|v_n(w_n) - \nabla_\theta J(\theta^{(n)})\|^2 + 2\|\nabla_\theta J(\theta^{(n)})\|^2. \tag{49}$$

Taking the expectation on (47) and rearranging the terms, we have:

$$\begin{aligned}&\left(\frac{1}{2}\alpha - L_J\alpha^2\right)\mathbb{E}\left[\|\nabla_\theta J(\theta^{(n)})\|^2\right]\\ &\quad\leq \mathbb{E}\left[J(\theta^{(n+1)})\right] - J(\theta^{(n)}) + \left(\frac{1}{2}\alpha + L_J\alpha^2\right)\mathbb{E}\left[\|v_n(w_n) - \nabla_\theta J(\theta^{(n)})\|^2\right], &(50)\end{aligned}$$

where the last term in the RHS should be carefully controlled. To this end, we adopt triangle inequality to attain the following inequality:

$$\begin{aligned}\|v_n(w_n) - \nabla_\theta J(\theta^{(n)})\|^2 &\leq 6\|v_n(w_n) - v_n(w_n^*)\|^2 + 6\|v_n(w_n^*) - h_n(w_n^*)\|^2\\ &\quad + 6\|h_n(w_n^*) - g(w_n^*, \theta^{(n)})\|^2 + 6\|g(w_n^*, \theta^{(n)}) - \nabla_\theta J(\theta^{(n)})\|^2. &(51)\end{aligned}$$

We can decompose the first three terms using the following fact: $\|x\|^2 = \sum_{k=1}^K \|x_k\|^2$ for any $x = [x_1, \ldots, x_K]^\top$.

Now, we are ready to control each term in (51). The first term in the RHS of (51) can be bounded as follows:

$$\left\|v_n(w_n) - v_n(w_n^*)\right\|^2 \tag{52}$$

$$= \sum_{k\in\mathcal{K}} \left\|v_{k,n}(w_{k,n}) - v_{k,n}(w_{k,n}^*)\right\|^2 \tag{53}$$

$$= \sum_{k\in\mathcal{K}} \left\|\tilde{\delta}_{k,n}(w_{k,n})\cdot\psi_{k,n} - \tilde{\delta}_{k,n}(w_{k,n}^*)\cdot\psi_{k,n}\right\|^2 \tag{54}$$

$$= \sum_{k\in\mathcal{K}} \left\|\left[(\tilde{\delta}_{k,n}(w_{k,n}) - \tilde{\delta}_{k,n}(w_{k,n}^*)\right]\cdot\psi_{k,n}\right\|^2 \tag{55}$$

$$\leq \sum_{k\in\mathcal{K}} \left\|\tilde{\delta}_{k,n}(w_{k,n}) - \tilde{\delta}_{k,n}(w_{k,n}^*)\right\|^2 \cdot \left\|\psi_{k,n}\right\|^2 \tag{56}$$

$$\leq \sum_{k\in\mathcal{K}} \Big\|(\tilde{r}_{k,n} + \gamma\phi^\top(\Psi_{k,n+1})w_{k,n} - \phi^\top(\Psi_{k,n})w_{k,n}) - (\tilde{r}_{k,n} + \gamma\phi^\top(\Psi_{k,n+1})w_{k,n}^* - \phi^\top(\Psi_{k,n})w_{k,n}^*)\Big\|^2 \tag{57}$$

$$= \sum_{k\in\mathcal{K}} \left\|\left[\gamma\phi^\top(\Psi_{k,n+1}) - \phi^\top(\Psi_{k,n})\right](w_{k,n} - w_{k,n}^*)\right\|^2 \tag{58}$$

$$\leq \sum_{k\in\mathcal{K}} \left\|\gamma\phi^\top(\Psi_{k,n+1}) - \phi^\top(\Psi_{k,n})\right\|^2 \cdot \left\|w_{k,n} - w_{k,n}^*\right\|^2 \tag{59}$$

$$\leq \sum_{k\in\mathcal{K}} (1+\gamma)^2 \|w_{k,n} - w_{k,n}^*\|^2 \tag{60}$$

$$= (1+\gamma)^2 \sum_{k\in\mathcal{K}} \|w_{k,n} - w_{k,n}^*\|^2 \tag{61}$$

where the second inequality is from Assumption 6, and the last inequality is due to Assumption 5.

The second term in the RHS of (51) can be bounded as follows:

$$\left\|v_n(w_n^*) - h_n(w_n^*)\right\|^2 \tag{62}$$

$$= \sum_{k\in\mathcal{K}} \left\|v_{k,n}(w_{k,n}^*) - h_{k,n}(w_{k,n}^*)\right\|^2 \tag{63}$$

$$= \sum_{k\in\mathcal{K}} \left\|\tilde{\delta}_{k,n}(w_{k,n}^*)\cdot\psi_{k,n} - \delta_{k,n}^*(w_{k,n}^*)\cdot\psi_{k,n}\right\|^2 \tag{64}$$

$$= \sum_{k\in\mathcal{K}} \left\|\left[\tilde{\delta}_{k,n}(w_{k,n}^*) - \delta_{k,n}^*(w_{k,n}^*)\right]\cdot\psi_{k,n}\right\|^2 \tag{65}$$

$$\leq \sum_{k\in\mathcal{K}} \left\|\tilde{\delta}_{k,n}(w_{k,n}^*) - \delta_{k,n}^*(w_{k,n}^*)\right\|^2 \cdot \left\|\psi_{k,n}\right\|^2 \tag{66}$$

$$\leq \sum_{k\in\mathcal{K}} \Big\|(\tilde{r}_{k,n} + \gamma\phi^\top(\Psi_{k,n+1})w_{k,n}^* - \phi^\top(\Psi_{k,n})w_{k,n}^*) - (\bar{r}_{k,n} + \gamma\phi^\top(\Psi_{k,n+1})w_{k,n}^* - \phi^\top(\Psi_{k,n})w_{k,n}^*)\Big\|^2 \tag{67}$$

$$= \sum_{k\in\mathcal{K}} \left|\left([\mathbf{C}^G]_k - \frac{1}{K}\mathbf{1}^\top\right)[r_{1,n},\cdots,r_{K,n}]^\top\right|^2 \tag{68}$$

$$\leq \sum_{k\in\mathcal{K}} \left\|[\mathbf{C}^G]_k - \frac{1}{K}\mathbf{1}^\top\right\|^2 \cdot \left\|[r_{1,n},\cdots,r_{K,t}]\right\|^2 \tag{69}$$

$$\leq \sum_{k\in\mathcal{K}} \left\|[\mathbf{C}^G]_k - \frac{1}{K}\mathbf{1}^\top\right\|^2 K r_{\max}^2 \tag{70}$$

$$\leq \sum_{k\in\mathcal{K}} 4K\left[(1+\nu^{-(K-1)})(1-\nu^{K-1})^G\right]^2 \cdot Kr_{\max}^2 \tag{71}$$

$$= 4K^3 r_{\max}^2\left((1+\nu^{-(K-1)})(1-\nu^{K-1})^G\right)^2, \tag{72}$$

where the second inequality is due to Assumption 6, and the last inequality is by the property of gossiping technique.

According to the definitions of $h_{k,n}$ and (43), the third term in the RHS of (51) can be written as follows:

$$\left\|h_n(w_n^*) - g(w_n^*, \theta^{(n)})\right\|^2 = \sum_{k\in\mathcal{K}} \left\|h_{k,n}(w_{k,n}^*) - g(w_{k,n}^*, \theta_k^{(n)})\right\|^2 \tag{73}$$

$$= \sum_{k\in\mathcal{K}} \left\|\delta_{k,n}^*(w_{k,n}^*)\cdot\psi_{k,n} - \mathbb{E}[\mathrm{Adv}_{w_{k,n}^*}(s_n, \boldsymbol{a}_n)\psi_{\theta_k^{(n)}}(s_n, a_{k,n})]\right\|^2. \tag{74}$$

Taking expectation over the filtration $\mathcal{F}_n$ on both sides of (74), we have:

$$\mathbb{E}\left[\left\|h_n(w_n^*) - g(w_n^*, \theta^{(n)})\right\|^2 | \mathcal{F}_n\right] \tag{75}$$

$$= \mathbb{E}\left[\sum_{k\in\mathcal{K}} \left\|h_{k,n}(w_{k,n}^*) - g(w_{k,n}^*, \theta_k^{(n)})\right\|^2 | \mathcal{F}_n\right] \tag{76}$$

$$= \mathbb{E}\left[\sum_{k\in\mathcal{K}} \left\|\delta_{k,n}^*(w_{k,n}^*)\psi_{k,n} - \mathbb{E}\big[\mathrm{Adv}_{w_{k,n}^*}(s_n, \boldsymbol{a}_n)\psi_{\theta_k^{(n)}}(s_n, a_{k,n})\big]\right\|^2 | \mathcal{F}_n\right] \tag{77}$$

$$= \mathbb{E}\left[\sum_{k\in\mathcal{K}} \left\|\delta_{k,n}^*(w_{k,n}^*)\psi_{k,n} - \mathbb{E}\big[\delta_{k,n}^*(w_{k,n}^*)\psi_{k,n}\big]\right\|^2 | \mathcal{F}_n\right] \tag{78}$$

$$= \mathbb{E}\left[\sum_{k\in\mathcal{K}} \left\|\big(\delta_{k,n}^*(w_{k,n}^*) - \mathbb{E}[\delta_{k,n}^*(w_{k,n}^*)]\big)\cdot\psi_{k,n}\right\|^2 | \mathcal{F}_n\right] \tag{79}$$

$$\leq \mathbb{E}\left[\sum_{k\in\mathcal{K}} \left|\delta_{k,n}^*(w_{k,n}^*) - \mathbb{E}[\delta_{k,n}^*(w_{k,n}^*)]\right|^2 \cdot \left\|\psi_{k,n}\right\|^2 | \mathcal{F}_n\right] \tag{80}$$

$$\leq \mathbb{E}\left[\sum_{k\in\mathcal{K}} \left|\delta_{k,n}^*(w_{k,n}^*) - \mathbb{E}[\delta_{k,n}^*(w_{k,n}^*)]\right|^2 | \mathcal{F}_n\right] \tag{81}$$

$$\leq \mathbb{E}\left[\sum_{k\in\mathcal{K}} \left|\delta_{k,n}^*(w_{k,n}^*)\right|^2 + \left|\mathbb{E}[\delta_{k,n}^*(w_{k,n}^*)]\right|^2 | \mathcal{F}_n\right] \tag{82}$$

$$= \sum_{k\in\mathcal{K}} \mathbb{E}\left[\left|\delta_{k,n}^*(w_{k,n}^*)\right|^2 + \left|\mathbb{E}[\delta_{k,n}^*(w_{k,n}^*)]\right|^2 | \mathcal{F}_n\right] \tag{83}$$

$$\leq \sum_{k\in\mathcal{K}} 2\mathbb{E}\left[\left|\delta_{k,n}^*(w_{k,n}^*)\right|^2 | \mathcal{F}_n\right] + 2\mathbb{E}\left[\left|\mathbb{E}[\delta_{k,n}^*(w_{k,n}^*)]\right|^2 | \mathcal{F}_n\right] \tag{84}$$

$$\leq \sum_{k\in\mathcal{K}} 4(r_{\max} + (1+\gamma)R_w)^2 \tag{85}$$

$$= 4N(r_{\max} + (1+\gamma)R_w)^2, \tag{86}$$

where the second inequality is from Assumption 6. The last inequality is due to

$$\|\delta_{k,n}^*(w_{k,n})\| = \|\bar{r}_{k,n} + \gamma\phi^\top(\Psi_{k,n+1})w_{k,n}^* - \phi^\top(\Psi_{k,n})w_{k,n}^*\| \tag{87}$$

$$= \|\bar{r}_{k,n} + [\gamma\phi^\top(\Psi_{k,n+1}) - \phi^\top(\Psi_{k,n})]w_{k,n}^*\| \tag{88}$$

$$\leq \|\bar{r}_{k,n}\| + \|\gamma\phi^\top(\Psi_{k,n+1}) - \phi^\top(\Psi_{k,n})\|\cdot\|w_{k,n}^*\| \tag{89}$$

$$\leq \|\bar{r}_{k,n}\| + \big(\|\gamma\phi^\top(\Psi_{k,n+1})\| + \|\phi^\top(\Psi_{k,n})\|\big)\cdot\|w_{k,n}^*\| \tag{90}$$

$$\leq r_{\max} + (\gamma + 1)\, R_w, \tag{91}$$

where the last inequality is due to Assumptions 1 and 5 as well as the 2-norm bound on the equilibrium point $w^*_{k,n}$ [34].

The last term in the RHS of (51) can be bounded as follows:

$$\left\|g(w_n^*, \theta^{(n)}) - \nabla_\theta J(\theta^{(n)})\right\|^2 \tag{92}$$
$$= \left\|\mathbb{E}\left[\text{Adv}_{w_n^*}(s_n, \boldsymbol{a}_n)\psi_{\theta^{(n)}}(s_n, \boldsymbol{a}_n)\right] - \mathbb{E}\left[\text{Adv}_{\theta^{(n)}}(s_n, \boldsymbol{a}_n)\psi_{\theta^{(n)}}(s_n, \boldsymbol{a}_n)\right]\right\|^2 \tag{93}$$
$$= \left\|\mathbb{E}\left[(\text{Adv}_{w_n^*}(s_n, \boldsymbol{a}_n) - \text{Adv}_{\theta^{(n)}}(s_n, \boldsymbol{a}_n))\psi_{\theta^{(n)}}(s_n, \boldsymbol{a}_n)\right]\right\|^2 \tag{94}$$
$$\leq \left(\mathbb{E}\big[\|(\text{Adv}_{w_n^*}(s_n, \boldsymbol{a}_n) - \text{Adv}_{\theta^{(n)}}(s_n, \boldsymbol{a}_n))\psi_{\theta^{(n)}}(s_n, \boldsymbol{a}_n)\|\big]\right)^2 \tag{95}$$
$$\leq \left(\mathbb{E}\left[|\text{Adv}_{w_n^*}(s_n, \boldsymbol{a}_n) - \text{Adv}_{\theta^{(n)}}(s_n, \boldsymbol{a}_n)| \cdot \|\psi_{\theta^{(n)}}(s_n, \boldsymbol{a}_n)\|\right]\right)^2 \tag{96}$$
$$\leq \left(\mathbb{E}\left[|\text{Adv}_{w_n^*}(s_n, \boldsymbol{a}_n) - \text{Adv}_{\theta^{(n)}}(s_n, \boldsymbol{a}_n)|\right]\right)^2 \tag{97}$$
$$= \left(\mathbb{E}\left[|\mathbb{E}[\gamma V_{w_n^*}(s_{n+1})|s_n, \boldsymbol{a}_n] - V_{w_n^*}(s_n) - \mathbb{E}[\gamma V_{\theta^{(n)*}}(s_{n+1})|s_n, \boldsymbol{a}_n] + V_{\theta^{(n)*}}(s_n)|\right]\right)^2 \tag{98}$$
$$\leq \left(\mathbb{E}\left[|\mathbb{E}[\gamma V_{w_n^*}(s_{n+1}) - \gamma V_{\theta^{(n)*}}(s_{n+1})|s_n, \boldsymbol{a}_n]| + |V_{w_n^*}(s_n) - V_{\theta^{(n)*}}(s_n)|\right]\right)^2 \tag{99}$$
$$\leq \left(\mathbb{E}\left[\mathbb{E}[|\gamma V_{w_n^*}(s_{n+1}) - \gamma V_{\theta^{(n)*}}(s_{n+1})||s_n, \boldsymbol{a}_n] + |V_{w_n^*}(s_n) - V_{\theta^{(n)*}}(s_n)|\right]\right)^2 \tag{100}$$
$$= \left(\mathbb{E}[|\gamma V_{w_n^*}(s_n) - \gamma V_{\theta^{(n)*}}(s_n)|] + \mathbb{E}\left[|V_{w_n^*}(s_n) - V_{\theta^{(n)*}}(s_n)|\right]\right)^2 \tag{101}$$
$$\leq (1+\gamma)^2\left(\mathbb{E}\left[|V_{w_n^*}(s_n) - V_{\theta^{(n)*}}(s_n)|\right]\right)^2 \tag{102}$$
$$\leq (1+\gamma)^2\mathbb{E}\left[|V_{w_n^*}(s_n) - V_{\theta^{(n)*}}(s_n)|^2\right] \tag{103}$$
$$\leq (1+\gamma)^2\xi_{\text{approx}}^{\text{critic}}, \tag{104}$$

where $\xi_{\text{approx}}^{\text{critic}}$ is the error bound on the linear approximation of value function.

Combining everything together, we can upper bound the RHS of (51) as

$$\mathbb{E}\left[\|v_n(w_n) - \nabla_\theta J(\theta^{(n)})\|^2\right] \tag{105}$$
$$\leq 6(1+\gamma)^2\sum_{k\in\mathcal{K}}\|w_{k,n} - w_{k,n}^*\|^2 + 24K^3 r_{\max}^2\left((1+\nu^{-(K-1)})(1-\nu^{K-1})^G\right)^2$$
$$+ 24K(r_{\max} + (1+\gamma)R_w)^2 + 6(1+\gamma)^2\xi_{\text{approx}}^{\text{critic}}. \tag{106}$$

Therefore, we have:

$$\left(\frac{1}{2}\alpha - L_J\alpha^2\right)\mathbb{E}\left[\|\nabla_\theta J(\theta^{(n)})\|^2\right]$$
$$\leq \mathbb{E}\left[J(\theta^{(n+1)})\right] - \mathbb{E}[J(\theta^{(n)})] + \left(\frac{1}{2}\alpha + L_J\alpha^2\right)\left(6(1+\gamma)^2\sum_{k\in\mathcal{K}}\|w_{k,n} - w_{k,n}^*\|^2\right. \tag{107}$$
$$\left. + 24K^3 r_{\max}^2\left((1+\nu^{-(K-1)})(1-\nu^{K-1})^G\right)^2 + 24K(r_{\max} + (1+\gamma)R_w)^2 + 6(1+\gamma)^2\xi_{\text{approx}}^{\text{critic}}\right). \tag{108}$$

By setting step-size $\alpha = \frac{1}{4L_J}$ and dividing both sides of previous equation by $\frac{1}{16L_J}$, we further obtain:

$$\mathbb{E}\left[\|\nabla_\theta J(\theta^{(n)})\|^2\right]$$
$$\leq 16L_J\mathbb{E}\left[J(\theta^{(n+1)})\right] - 16L_J\mathbb{E}[J(\theta^{(n)})] + 18(1+\gamma)^2\sum_{k\in\mathcal{K}}\|w_{k,n} - w_{k,n}^*\|^2$$

$$+ 72K^3 r_{\max}^2 \left((1+\nu^{-(K-1)})(1-\nu^{K-1})^G\right)^2 + 72K(r_{\max}+(1+\gamma)R_w)^2 + 18(1+\gamma)^2\xi_{\text{approx}}^{\text{critic}}. \tag{109}$$

Let $\hat{N}$ be a random integer variable uniformly taken from $(1, N)$. If we take summation over $n = \{1, \ldots, N\}$ and divide it by N, we have

$$\mathbb{E}\left[\|\nabla_\theta J(\theta^{(\hat{N})})\|^2\right] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}[\|\nabla_\theta J(\theta^{(n)})\|^2] \tag{110}$$

$$\leq \frac{16L_J(\mathbb{E}\left[J(\theta^{(n)})\right] - \mathbb{E}\left[J(\theta^{(0)})\right])}{N} + 18(1+\gamma)^2\frac{\sum_{n=1}^{N}\sum_{k\in\mathcal{K}}\|w_{k,n}-w_{k,n}^*\|^2}{N}$$
$$+ 72K^3 r_{\max}^2 \left((1+\nu^{-(K-1)})(1-\nu^{K-1})^G\right)^2 + 72K(r_{\max}+(1+\gamma)R_w)^2 + 18(1+\gamma)^2\xi_{\text{approx}}^{\text{critic}} \tag{111}$$

$$\leq \frac{16L_J\mathbb{E}\left[J(\theta^{(n)})\right]}{N} + 18(1+\gamma)^2\frac{\sum_{n=1}^{N}\sum_{k\in\mathcal{K}}\|w_{k,n}-w_{k,n}^*\|^2}{N}$$
$$+ 72K^3 r_{\max}^2 \left((1+\nu^{-(K-1)})(1-\nu^{K-1})^G\right)^2 + 72K(r_{\max}+(1+\gamma)R_w)^2 + 18(1+\gamma)^2\xi_{\text{approx}}^{\text{critic}} \tag{112}$$

$$\leq \frac{16L_J r_{\max}}{N(1-\gamma)} + 18(1+\gamma)^2\frac{\sum_{n=1}^{N}\sum_{k\in\mathcal{K}}\|w_{k,n}-w_{k,n}^*\|^2}{N}$$
$$+ 72K^3 r_{\max}^2 \left((1+\nu^{-(K-1)})(1-\nu^{K-1})^G\right)^2 + 72K(r_{\max}+(1+\gamma)R_w)^2 + 18(1+\gamma)^2\xi_{\text{approx}}^{\text{critic}}. \tag{113}$$