Meta-Reinforcement Learning in Time-Varying UAV Communications: Adaptive Anti-Jamming Channel Selection

Linzi HU [1], Yumeng SHAO [1], Yuwen QIAN [1], Feng DU [1], Jun LI [1], Yan LIN [1,3], Zhe WANG [2]

[1] School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[2] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
[3] National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China

{linzihu, shaoyumeng, admon, chary.df, jun.li, yanlin, zwang}@njust.edu.cn

Submitted March 28, 2024 / Accepted May 31, 2024 / Online first June 17, 2024

Abstract. *Unmanned Aerial Vehicle (UAV) communication networks are vulnerable to malicious jamming and co-channel interference, deteriorating the performance of the networks. Therefore, the exploration of anti-jamming methods to enhance communication security becomes a significant challenge. In this paper, we propose a novel anti-jamming channel selection scheme in a multi-channel multi-UAV network. We first formulate the anti-jamming problem as a Partially Observable Stochastic Game (POSG), where the UAV pairs with partial observability compete for a limited number of communication channels against a Markov jammer. To ensure rapid adaptation to the dynamic jamming environment, we propose a Meta-Mean-Field Q-learning (MMFQ) algorithm, which provides a Nash Equilibrium (NE) solution to the POSG problem. Furthermore, we derive the expressions of the upper bound for the loss function of MMFQ and prove the convergence of the proposed algorithm. Simulation results demonstrate that the proposed algorithm can achieve a superior average reward compared to the benchmark algorithms, facilitating throughput enhancement and resource utilization increase, especially for large-scale UAV communication networks.*

Keywords

Unmanned aerial vehicle (UAV) communication, anti-jamming, meta-reinforcement learning, mean field

1. Introduction

Unmanned Aerial Vehicles (UAVs) have been recently applied to various scenarios to improve operational efficiency, enhance situational awareness, reduce risks, and increase capabilities [1–3]. However, UAV communication links are vulnerable to co-channel interference and malicious jamming. In this case, a practical and reliable communication method is essential for real-time data transmission and security insurance in multi-UAV scenarios. Therefore, anti-jamming methods are employed in UAV communication systems to mitigate the effects of malicious jamming. The traditional anti-jamming methods, such as Frequency Hopping Spread Spectrum (FHSS) [4], adaptive beamforming [5], adaptive power control [6], and Direct Sequence Spread Spectrum (DSSS) [7], typically require high bandwidth and may not provide multi-UAV coordination.

To address the above challenges, Reinforcement Learning (RL) has been adopted in the field of anti-jamming communication. The UAVs can adopt model-free RL techniques to optimize their real-time anti-jamming policies through interaction with complex and unknown environments. For example, the authors of [8] proposed a Sequential Deep Reinforcement Learning Algorithm (SDRLA) that lacks prior information within a jamming environment. However, the computer vision used in SDRLA to acquire insights and characteristics about jamming patterns causes an infinite state of spectrum waterfall, which requires significant computational resources. To address the problem, an improved deep RL-based anti-jamming algorithm was proposed in [9], which substitutes infinite states with the spectrum differences between adjacent time slots. In addition, a collaborative multi-agent RL-based anti-jamming algorithm was proposed in [10] to optimize the quality of service by jointly optimizing the channel and power allocation for UAVs. Similarly, reference [11] designed a knowledge-based RL method to deal with the problem of high-dimensional state space in UAVs against intelligent jamming. Furthermore, authors in [12] proposed a multi-agent anti-jamming RL-based method with partially overlapping channels, i.e., some channels use adjacent frequency ranges. There exists a trade-off between the spectrum utilization efficiency and the anti-jamming performance, not only increasing the number of available channels but also exacerbating more serious interference due to channel overlap. The aforementioned RL-based approaches tend to optimize the anti-jamming policy against the stationary

DOI: 10.13164/re.2024.0417

environment and jamming policy, where the anti-jamming policy needs to be retrained to adapt to the new task when the environment or jamming policy changes.

Meta-Reinforcement Learning (Meta-RL) models can reduce the number of samples required to learn a new task, such as a new jamming policy or new location of the jammers, by utilizing the prior knowledge of the concerned tasks. This ability to learn across tasks allows the model to quickly adapt to new jamming environments after a small amount of trial and error. As a response to the constraints inherent in existing RL-based studies, Meta-RL has been incorporated into communication systems. For example, reference [13] leveraged Meta-RL to learn the model for digital twin in industrial Internet of Things (IoT), which improves the generalization and fast adaptation of the model to a new environment. Similarly, a meta-based deep RL algorithm was proposed in [14] to enhance the fast adaptability of the resource allocation policy for the dynamic vehicle to everything communications. Moreover, Meta-RL was adopted in such as task offloading of mobile-edge computing environments [15], real-time scheduling wireless traffic [16], and load balancing for multi-band downlink cellular networks [17], improving the model adaptability across different environments or objectives. Despite the advancement of Meta-RL in communication systems, this technique has not been studied for anti-jamming communication, especially for multi-UAV scenarios, failing to consider the relation among different jamming environments.

In this paper, we propose an adaptive anti-jamming channel selection algorithm for a self-organized UAV communication network with a massive number of self-interested and non-cooperative UAV users. Considering the large-scale scenario, mean-field theory [18] offers an effective mathematical tool to approximate the aggregate behavior of a large number of strategic agents. In UAV communication systems, mean-field has been applied to approximate the aggregate behaviors of the wireless agents [20], [21]. In addition, a mean-field-based anti-jamming method was proposed in [22] to achieve the Nash equilibrium solution of the Markov game in a large-scale Internet of Things network.

The main contributions of this paper are listed as follows.

- We consider a UAV anti-jamming network, where the jammers intend to attack the UAVs' communication while the UAV pairs fight for the scarce spectrum access opportunity over several channels. Then, the anti-jamming problem is formulated as a Partially Observable Stochastic Game (POSG), where each UAV pair can only observe a portion of the network environment.

- We develop the mean-field game for the multi-agent scenario to further streamline the anti-jamming problem with numerous UAV pairs, where each UAV pair simply interacts with the interference from an average neighbor and the malicious jammer.

- We propose a Meta-RL-based algorithm, named Meta-Mean-Field Q-learning (MMFQ) to solve the adaptive channel selection under varying jamming environments.

- We prove the convergence and the fast adaptation performance of the proposed algorithm. Specifically, we derive that the objective function of MMFQ is bounded in the presence of task variability and the convergence errors decrease as task similarity increases. We further evaluate the distance between the output generated by RL and the optimized model of adaptation theoretically, which indicates that greater task similarity improves the performance of the MMFQ algorithm. The simulation results demonstrate that compared to the benchmark algorithms, the proposed algorithm converges more quickly and achieves a greater throughput.

The remainder of the paper is structured as follows. The system model is introduced in Sec. 2. We formulate the channel competition among the UAV pairs as a POSG in Sec. 3. Section 4 describes the mean field Q-learning method for multi-UAV scenarios. Section 5 describes the Meta-RL algorithm MMFQ for multiple tasks. Section 6 presents the convergence analysis and fast adaptation performance evaluation of MMFQ. Section 7 shows the experimental results. Finally, we conclude our work in Sec. 8.

2. System Model

Figure 1 shows the proposed UAV communication network, where a UAV pair consists of a UAV transmitter and a UAV receiver. Each UAV pair competes for the available spectrum resources while avoiding interference from both the jammers and other UAV pairs. Denote the set of UAV pairs as $\mathcal{N} = \{1, \cdots, n, \cdots, N\}$, the set of channels as $\mathcal{M} = \{1, \cdots, m, \cdots, M\}$, and the set of jammers as $\mathcal{J} = \{1, \cdots, j, \cdots, J\}$.

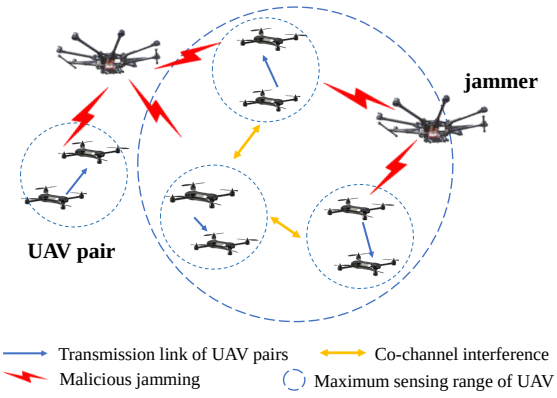

Fig. 1. The UAV anti-jamming network with N UAV pairs and J malicious jammers, where each UAV pair consists of a UAV transmitter and a UAV receiver. For a typical UAV receiver, it is subject to malicious interference from jammers and co-channel interference from other UAV transmitters.

For the n-th UAV pair, denote the transmit power of the UAV transmitter as P_n, and the channel selection action in t-th time slot as $a_n(t) \in \mathcal{M}$. The transmit power of jamming signals at the j-th jammer is P_j, and its channel selection action is denoted as $g_j(t) \in \mathcal{M}$, which operates according to an independent Markov process, i.e., the jammer follows a state transition matrix to determine which channel to select the next time.

We assume the scenario where the positions of UAVs are time-varying following a Gaussian Markov model [23], and the receivers are restricted to a certain range of their corresponding transmitters. Therefore, the moving velocity $v_n(t)$ and direction $\theta_n(t)$ of the UAV transmitter n in t-th time slot is respectively expressed as

$$v_n(t) = k_1 v_n(t-1) + (1-k_1)\,\bar{v}_n + \sqrt{1-k_1^2}\varphi_n \tag{1}$$

and

$$\theta_n(t) = k_2 \theta_n(t-1) + (1-k_2)\,\bar{\theta}_n + \sqrt{1-k_2^2}\psi_n \tag{2}$$

where $0 \le k_1, k_2 \le 1$ denote the memory coefficient of the velocity and direction, respectively. $v_n(t)$ is related to average velocity $\bar{v}_n$ and the randomness of the velocity φ_n. $\bar{\theta}_n$ and ψ_n denote the average direction and the randomness of the direction, respectively. φ_n and ψ_n follow Gaussian distributions.

As shown in Fig. 2, we consider a time-slotted framework for the UAV communication network. All the UAVs update their locations according to (1) and (2) at the start of each time slot, and all the UAV transmitter performs wideband spectrum sensing across all accessible channels. Following that, the UAV transmitter determines the channel access decision, utilizing the historical information of the previous time slots and the present spectrum sensing observation. After that, each UAV transmitter transmits data on the selected channel. Ultimately, each UAV transmitter broadcasts its channel selection to its neighbors in the present time slot and receives an acknowledgment (ACK) or non-acknowledgment (NACK) signal from its UAV receiver. Furthermore, the jammer network is also time-slotted, but not necessarily synchronized with that of the UAV network.

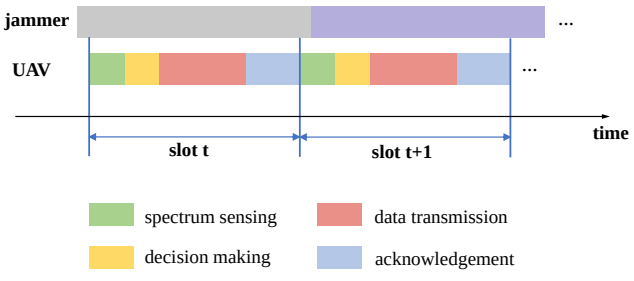

Fig. 2. Transmission slot structure of UAV user and jammer.

For a typical UAV pair, besides the malicious interference from the jammers, there exists co-channel interference primarily depending on the distance between the typical UAV pair and the interfering user pair. We define the neighbor of the n-th UAV pair as that if the interference at the n-th UAV receiver from the i-th UAV transmitter exceeds a pre-defined threshold I_{th}, i.e.,

$$\mathbb{E}\left[P_i h_{i,n}(t) d_{i,n}^{-\lambda}\right] \ge I_{\text{th}} \tag{3}$$

where $h_{i,n}(t)$ represents the channel fading gain, $d_{i,n}$ denotes the distance between the i-th UAV transmitter and the n-th UAV receiver, and λ is the path loss coefficient. We further define the neighbor set of the n-th UAV pair as

$$\mathcal{N}(n) = \left\{ i | i \in \mathcal{N} \backslash n, d_{i,n} \le d_{\text{th}} \right\} \tag{4}$$

where $d_{\text{th}} = \left(p_0 \mathbb{E}\left[h_{i,n}(t)\right] / I_{\text{th}}\right)^{\frac{1}{\lambda}}$ is the neighbor distance threshold with P_i fixed as p_0.

The signal-to-interference plus noise ratio (SINR) threshold at the UAV receiver is denoted as χ, and if the actual channel capacity at the UAV receiver is larger than the threshold, the reception of the signal is successful and an ACK is returned to the UAV transmitter. Then, the achievable throughput at the n-th UAV pair in t-th time slot is expressed as

$$D_n(t) = \begin{cases} B\tau_{\text{tr}} \log_2(1+\chi), & \text{if SINR} \ge \chi \\ 0, & \text{if SINR} < \chi \end{cases} \tag{5}$$

with

$$\text{SINR} = \frac{P_n d_{n,n}^{-\lambda}}{I_n(t) + J_n(t) + N_n} \tag{6}$$

where B denotes the channel bandwidth, τ_{tr} is the duration of data transmission, $d_{n,n}$ represents the distance between the transmitter and receiver for the n-th UAV pair, and N_n is the noise power. In addition, $I_n(t)$ and $J_n(t)$ respectively denote the interference received from all other UAV transmitters and jammers, which can be given by

$$I_n(t) = \sum_{i \in \mathcal{N} \backslash n} P_i h_{i,n}(t) d_{i,n}^{-\lambda}(t) f\left(a_i(t), a_n(t)\right) \tag{7}$$

and

$$J_n(t) = \sum_{j \in \mathcal{J}} P_j h_{j,n}(t) d_{j,n}^{-\lambda}(t) f\left(g_j(t), a_n(t)\right) \tag{8}$$

where $d_{i,n}$ denotes the distance between the i-th UAV transmitter and the n-th UAV receiver,$d_{j,n}$ denotes the distance between the j-th jammer and the n-th UAV receiver, and $f(\cdot)$ is an indicator function that describes the co-channel event of the two nodes of x and y, i.e.,

$$f(x,y) = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{if } x \ne y \end{cases}. \tag{9}$$

For each UAV pair n, it aims to optimize its channel selection that maximizes the long-term expected achievable rate, i.e.,

$$\max_{a_n(t) \in \mathcal{M}, \forall t} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t D_n(t)\right] \tag{10}$$

where $\gamma \in [0,1)$ is a discount factor, meaning the immediate achievable rate is more important than the future.

3. Non-Cooperative UAV Network in Anti-Jamming Game

In this paper, we formulate the anti-jamming problem as a POSG, where self-interested each UAV pair has partial observation of the jammers' channels and locations. POSG can be modeled as the tuple $\Gamma \triangleq (\mathcal{N}, S, A, O, P, \mathcal{R})$: where $\mathcal{N}$ is the set of UAV pairs, S denotes the state space, $A = A^1 \times \ldots \times A^N$ is the joint action set of all UAV pairs, and $O = O^1 \times \ldots \times O^N$ is the joint observations set of all UAV pairs. $P : S \times A \times S \to [0, 1]$ denotes the unknown transition probability between states after actions, and $\mathcal{R} : S \times A \to \mathbb{R}$ is the reward function. We define the corresponding elements in detail as follows.

- State s_t: The condition of the overall network environment, including the positions of the UAVs and jammers, and the jammers' selected channels in the previous time slot.

- Observation $\mathbf{o}_t^n$: The information observed by the n-th UAV pair in t-th time slot, denoted as $\mathbf{o}_t^n = \left[F_{t-1}^n, C_{t-1}^n, G_t^{j,n}\right]$ where F_{t-1}^n is an ACK/NACK signal, indicating the previous transmission is success or outage, C_{t-1}^n is the jammed channels sensed in the previous time slot, and $G_t^{j,n}$ is the distance between the UAV pair and the jammer. Specifically, $F_{t-1}^n \in \{0, 1\}$ where 0 signifies the communication failed for the n-th UAV pair, i.e., the SINR at the UAV receiver is lower than the threshold χ, and 1 indicates successful communication. Moreover, it holds that $C_{t-1}^n = \left[j_{t-1}^n(1), \ldots, j_{t-1}^n(m), \ldots, j_{t-1}^n(M)\right]$. If the channel m is jammed by the jammer, $j_t^n(m)$ is set to 1, and 0 otherwise. In addition, $G_t^{j,n} = \left[d_t^{j_1,n_t}, d_t^{j_2,n_t}, d_t^{j_1,n_r} d_t^{j_2,n_r}\right]$ where $d_t^{j_1,n_t}$ and $d_t^{j_2,n_t}$ denote the distance between the jammers and the UAV transmitter, $d_t^{j_1,n_r}$ and $d_t^{j_2,n_r}$ denote the distance between the jammers and the UAV receiver. The joint observation of the UAV network is denoted as $\mathbf{o}_t = \left[\mathbf{o}_t^1, \mathbf{o}_t^2, \ldots, \mathbf{o}_t^N\right]$.

- Action a_t^n: The channel selected by UAV pair n in t-th time slot, $a_t^n \in \mathcal{A}^n = \{1, \ldots, M\}$. The joint action set of the UAV swarm is defined as $\mathbf{a}_t = \left[a_t^1, a_t^2, \ldots, a_t^N\right]$.

- Reward r_t^n: The reward received by the n-th UAV pair in t-th time slot, is defined as $r_t^n = \sum_{t=0}^{t} \gamma^t D_n(t)$, where $D_n(t)$ was given in (5) and γ is the discount factor same as in (10). The system reward is defined as $\mathbf{r}_t = \left[r_t^1, r_t^2, \ldots, r_t^N\right]$.

Denote $\pi^n : O \to \Omega(A^n)$ as the channel selection policy for the n-th UAV pair, where $\Omega(A^n)$ is the set of probability distributions on the action space A^n. $\boldsymbol{\pi}_t \triangleq \left[\pi_t^1, \ldots, \pi_t^N\right]$ is the joint policy of all UAV pairs in t-th time slot. Each UAV pair learns to find an optimal policy $\left(\pi_t^n\right)^*$ to maximize its own long-term expected reward. Due to the co-channel interference, the joint policy $\boldsymbol{\pi}_t$ of all UAV pairs determines the optimization of long-term expected reward for the n-th UAV pair. Therefore, the value function V of the n-th UAV pair can be defined as the expected cumulative discounted reward under the joint policy $\boldsymbol{\pi}_t$, i.e.,

$$V_{\boldsymbol{\pi}_t}^n\left(\mathbf{o}_t^n\right) = V^n\left(\mathbf{o}_t^n, \boldsymbol{\pi}_t\right) = \sum_{t=0}^{T} \gamma^t \mathbb{E}_{\boldsymbol{\pi}_t, p}\left[r_t^n | o_0^n = o, \pi_t\right] \quad (11)$$

where o is the initial observation, p is the transition probability, and $\gamma \in [0, 1)$ is a discount factor. The joint observation $\mathbf{o}_t$ is commonly assumed as a mapping of the environment states s. Alternatively, the observation $\mathbf{o}_t$ obtained in state s follows a certain probability distribution. Therefore, the actual environmental state s conforms to a posterior distribution based on the observed $\mathbf{o}_t^n$ when all UAV pairs obtain a joint observation $\mathbf{o}_t$. Then, the objective function of each UAV pair is expressed as

$$\left(\pi_t^n\right)^* = \arg\max_{\pi_t^n} V^n\left(\mathbf{o}_t^n, \pi_t^n\right). \quad (12)$$

The NE of the POSG can be defined as the joint policy $\boldsymbol{\pi}_t^* \triangleq \left[\left(\pi_t^1\right)^*, \ldots, \left(\pi_t^n\right)^*\right]$, such that for all $o \in O$, $n \in 1, \cdots, N$ and all valid π_t^n, it satisfies with

$$V_{\boldsymbol{\pi}_t^*}^n\left(\mathbf{o}_t^n\right) = V_{\left(\pi_t^n\right)^*, \left(\boldsymbol{\pi}_t^{-n}\right)^*}^n\left(\mathbf{o}_t^n\right) \geq V_{\pi_t^n, \left(\boldsymbol{\pi}_t^{-n}\right)^*}^n\left(\mathbf{o}_t^n\right) \quad (13)$$

where $\left(\boldsymbol{\pi}_t^{-n}\right)^* \triangleq \left[\left(\pi_t^1\right)^*, \ldots, \left(\pi_t^{n-1}\right)^*, \left(\pi_t^{n+1}\right)^*, \ldots, \left(\pi_t^N\right)^*\right]$ represents the joint policy of all UAV pairs except the n-th UAV pair. For a POSG of N UAV pairs, there is at least one stable NE solution [24].

4. Mean Field Q-Learning for Solving the Large-Scale Anti-Jamming Game

In the context of a large-scale non-cooperative UAV communication network, addressing the Nash equilibrium of the POSG poses significant challenges due to its high computational complexity. To mitigate this issue, we employ mean-field game theory to streamline the POSG framework. Each UAV pair interacts with the collective interference generated by the entire population of jammers and other UAV pairs, integrating individual behaviors with aggregate effects. Subsequently, the population dynamics are updated based on the channel access policies learned by each UAV pair.

According to the value function defined in (11), the Q-value function ($Q_\pi^n : S \times A^1 \times \cdots \times A^N \to \mathbb{R}$) of the n-th UAV pair under the joint policy $\boldsymbol{\pi}_t$ can be given in

$$Q^n\left(\mathbf{o}_t^n, \mathbf{a}_t\right) = r_t^n\left(\mathbf{o}_t^n, \mathbf{a}_t\right) + \gamma \mathbb{E}_p\left[V_{\boldsymbol{\pi}_t}^n\left(o_{t+1}^n\right)\right] \quad (14)$$

where o_{t+1}^n denotes the observation in the next time slot. We can express the value function of local observation $\mathbf{o}_t^n$ in terms of the Q-value function as

$$V_{\boldsymbol{\pi}_t}^n\left(\mathbf{o}_t^n\right)=\mathbb{E}_{\mathbf{a}_t\sim\boldsymbol{\pi}_t}\left[Q^n\left(\mathbf{o}_t^n,\mathbf{a}_t\right)\right]. \tag{15}$$

The dimension of the joint actions $\mathbf{a}_t$ scales linearly with the number of UAV pairs, significantly increasing computational complexity within large-scale networks. Furthermore, learning the Q-function $Q^n\left(\mathbf{o}_t^n,\mathbf{a}_t\right)$ becomes impractical for individual UAV pairs due to the absence of information regarding the other non-cooperative UAV pairs. To address this challenge, we propose a decomposition strategy for the Q-function by using pairwise local interactions among neighboring UAV pairs, i.e.,

$$Q^n\left(\mathbf{o}_t^n,\mathbf{a}_t\right)=\frac{1}{N^n}\sum_{i\in\mathcal{N}(n)}Q^n\left(\mathbf{o}_t^n,a_t^n,a_t^i\right) \tag{16}$$

where $\mathcal{N}(n)$ is the neighboring UAV transmitters set for the n-th UAV pair, given in (4), $N^n=|\mathcal{N}(n)|$ is the number of neighbors.

The action of the n-th UAV pair a_t^n is a discrete variable representing the index of channels as stated in Sec. 2. Therefore, we denote the one-hot action a_t^i of each neighbor as

$$a_t^i=\bar{a}_t^n+\delta a_t^{n,i} \tag{17}$$

where $\bar{a}_t^n$ is the average action of all the neighbors for the n-th UAV pair, given by $\bar{a}_t^n=\frac{1}{N^n}\sum_i a_t^i$, and $\delta a_t^{n,i}$ is a small fluctuation.

Assume that $Q^n\left(\mathbf{o}_t^n,a_t^n,a_t^i\right)$ is second-order derivable with respect to the action a_t^i, we perform the Taylor expansion for (16) as follows

$$\begin{aligned}
Q^n\left(o^n,\mathbf{a}\right)&=\frac{1}{N^n}\sum_i Q^n\left(o^n,a^n,a^i\right)\\
&=\frac{1}{N^n}\sum_i\Big[Q^n\left(o^n,a^n,\bar{a}^{\,n}\right)\\
&\quad+\nabla_{\bar{a}^{\,n}}Q^n\left(o^n,a^n,\bar{a}^{\,n}\right)\delta a^{n,i}\\
&\quad+\frac{1}{2}\delta a^{n,i}\nabla^2_{\bar{a}^{\,n,i}}Q^n\left(o^n,a^n,\tilde{a}^{n,i}\right)\delta a^{n,i}\Big]\\
&=Q^n\left(o^n,a^n,\bar{a}^{\,n}\right)\\
&\quad+\nabla_{\bar{a}^{\,n}}Q^n\left(o^n,a^n,\bar{a}^{\,n}\right)\left[\frac{1}{N^n}\sum_i\delta a^{n,i}\right]\\
&\quad+\frac{1}{2N^n}\sum_i\left[\delta a^{n,i}\nabla^2_{\tilde{a}^{\,n,i}}Q^n\left(o^n,a^n,\tilde{a}^{n,i}\right)\delta a^{n,i}\right]
\end{aligned} \tag{18}$$

$$=Q^n\left(o^n,a^n,\bar{a}^{\,n}\right)+\frac{1}{2N^n}\sum_i R^n_{o^n,a^n}\left(a^i\right)$$

$$\approx Q^n\left(o^n,a^n,\bar{a}^{\,n}\right) \tag{19}$$

where $R^n_{o^n,a^n}\left(a^i\right)=\delta a^{n,i}\cdot\nabla^2_{a^{n,i}}Q^n\left(o^n,a^n,\tilde{a}^{n,i}\right)\cdot\delta a^{n,i}$, represents the remainder of the Taylor polynomial with $\tilde{a}^{n,i}=\bar{a}^{\,n}+\sigma^{n,i}\delta a^{n,i}$ and $\sigma^{n,i}\in[0,1]$. According to (17), $\sum_i\delta a^{n,i}=0$ in (18), and thus the second term is eliminated. In addition, it can be proved that $R^n_{o^n,a^n}\left(a^i\right)$ is a small fluctuation near zero when $Q^n\left(\mathbf{o}_t^n,a_t^n,a_t^i\right)$ is a linear function. Therefore, the Taylor expansion can be approximated as the first term in (18).

With the mean-field approximation, the pairwise interaction $Q^n\left(o^n,a^n,a^i\right)$ among the n-th UAV pair and its all neighbors is streamlined to the interaction between the n-th UAV pair and a virtual UAV pair, where the virtual UAV pair is an average effect of all neighbors of the n-th UAV pair. And Equation (19) is defined as the mean-field Q-function.

5. Meta-Reinforcement Learning for Fast Adaptation to Multiple Tasks

We consider the jamming environment, i.e., the location of the jammer, changing in a three-dimensional grid space, which represents different tasks for the UAV anti-jamming communication. Conventional RL primarily aims at training agents to obtain an optimal policy for addressing a single predefined task, i.e., the fixed location of jammers for the proposed system, which encounters limitations when it comes to effectively dealing with multiple similar tasks. Meta-RL, which applies meta-learning to RL problems, offers agents a solution to the fast adaptation to the new tasks. By leveraging the power of Meta-RL and employing algorithms like Model-Agnostic Meta-Learning (MAML) [25], the agents can efficiently attain and transfer knowledge among numerous interconnected tasks, improving their adaptability and performance in various environments. Therefore, the UAV pairs can quickly adapt to the new jamming environment and obtain the optimal policy with the Meta-RL-based algorithm when the jammers' locations change.

In this paper, we propose the MMFQ framework, which is a Meta-RL-based algorithm that combines meta-learning and mean-field Q-learning. As shown in Fig. 3, MMFQ consists of two distinctive phases: the training phase and the adaptation phase. The training phase aims to train the model parameters to obtain the optimal anti-jamming policy with mean-field Q-learning across sampled tasks. In the adaptation phase, the model intends to adapt to a new jamming environment with quick parameter adjustment based on the model parameters obtained during the training phase. MMFQ offers a model that undergoes fine-tuning when introduced to novel tasks through a gradient-based mean-field Q-learning approach. The objective of MMFQ lies in identifying model parameters that are responsive to variations in these tasks, thereby enabling minor parameter adjustments to yield significant enhancements. Therefore, we present the initial model by parameter vector θ and the optimal model trained by a meta-learner with parameter vector θ^*.

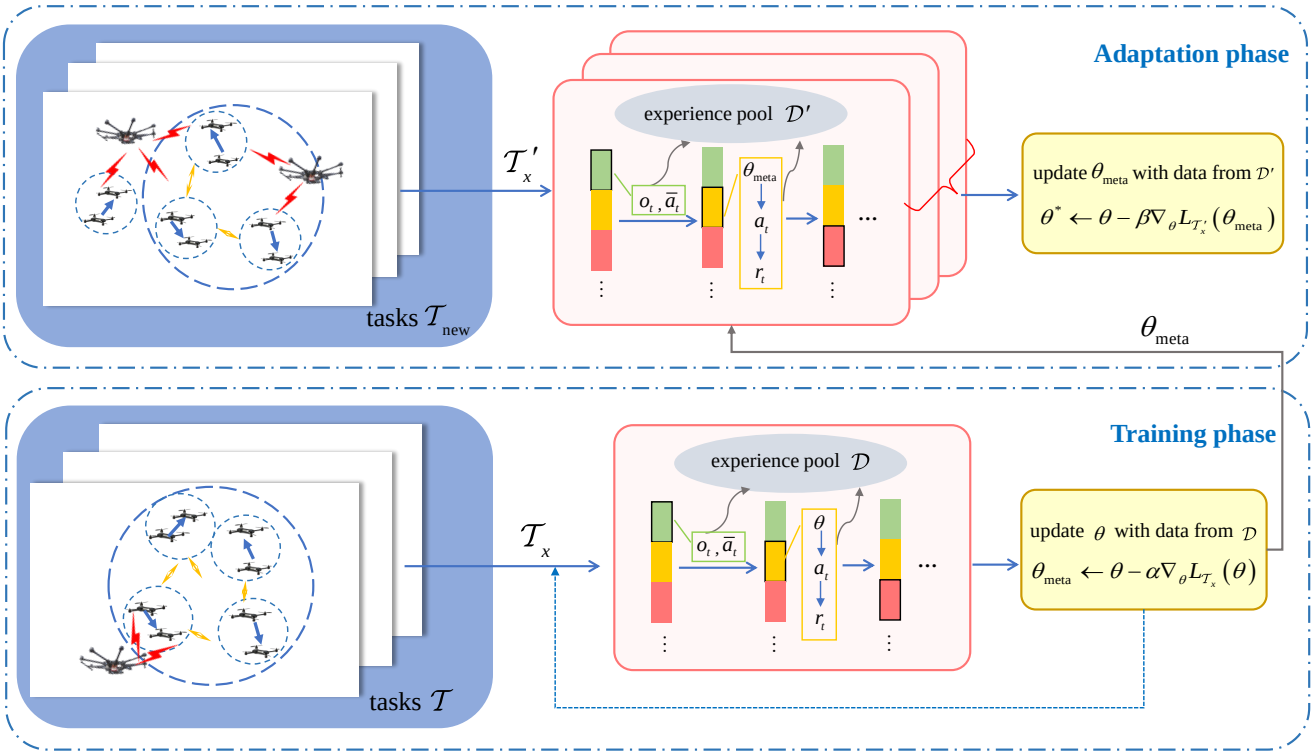

Fig. 3. MMFQ framework with a training phase and an adaptation phase, where users learn the meta-model through different training tasks in the training phase and update the model parameters during the adaptation phase with new tasks.

When implementing mean-field with the model θ in the MMFQ framework, Equation (19) can be represented as $Q^n\left(\theta_t^n; \mathbf{o}_t^n, a_t^n, \bar{a}_t^n\right)$, where θ_t^n denotes the model parameter of the UAV pair. Then, the UAV pair obtains a policy according to the output of the model, i.e.,

$$\pi_t^n(a_t^n|\mathbf{o}_t^n, \bar{a}_t^n) = \frac{\exp\left(-bQ^n\left(\theta_t^n; \mathbf{o}_t^n, a_t^n, \bar{a}_t^n\right)\right)}{\sum\limits_{a_t^{n'} \in \mathcal{A}^n} \exp\left(-bQ^n\left(\theta_t^n; \mathbf{o}_t^n, a_t^{n'}, \bar{a}_t^n\right)\right)} \quad (20)$$

where b is the Boltzmann constant.

Denote the maximum Q-value output by the model as $\hat{Q}^n\left(\theta_t^n; \mathbf{o}_t^n, a_t^n, \bar{a}_t^n\right)$. The UAV pair n is trained by minimizing the following loss function, given by

$$L\left(\theta_t^n\right) = \left(r_t^n + \gamma \max_{a_t^n} \hat{Q}^n\left(\theta_t^n; \mathbf{o}_t^n, a_t^n, \bar{a}_t^n\right) - Q^n\left(\theta_t^n; \mathbf{o}_t^n, a_t^n, \bar{a}_t^n\right)\right)^2. \quad (21)$$

As described in Algorithm 1, MMFQ consists of a training phase and an adaptation phase. In the training phase, we sample different tasks $\mathcal{T} \sim p(\tau)$, i.e., different locations of jammers, and the jammers are distributed in a three-dimensional gridded space with the set of $p(\tau)$. We suppose a task $\mathcal{T}_x \in \mathcal{T}$ is extracted for each episode, and there can be task repetition for different episodes. First, each UAV transmitter obtains the current observation $\mathbf{o}_t^n$ and the average action $\bar{a}_n$, and selects a channel a_t^n underlying θ_t^n according to (20). Then, each UAV pair communicates on the selected channel to acquire the reward r_t^n and a new observation $\mathbf{o}_{t+1}^n$. Therefore, a new average action $\bar{a}_t^n$ can be calculated. Subsequently, each UAV stores the array $\left\langle \mathbf{o}_t^n, a_t^n, r_t^n, \mathbf{o}_{t+1}^n, \bar{a}_t^n \right\rangle$ to the experience pool until the pool is full. Next, we update model parameters with the collected data. We draw a small batch of arrays denoted as $\mathcal{D}$, and evaluate $\nabla_\theta L_{\mathcal{T}_x}(\theta)$ using $\mathcal{D}$ with (21). Following this, the initial model θ is updated into meta model θ_{meta} as

$$\theta_{\text{meta}} \leftarrow \theta - \alpha\nabla_\theta L_{\mathcal{T}_x}(\theta) \quad (22)$$

where α is the learning rate.

In the adaptation phase, we first extract new tasks as $\mathcal{T}_{\text{new}} \sim p(\tau)$, which differs from the encountered tasks in the training phase. Similarly, each UAV pair obtains $\left\langle \mathbf{o}_t^n, a_t^n, r_t^n, o_{t+1}^n, \bar{a}_t^n \right\rangle$ in a manner parallel to the training phase, and then selects the communication channel a_t^n using θ_{meta}. We denote the new batch of arrays as $\mathcal{D}'$. Then, we evaluate $\nabla_\theta L_{\mathcal{T}_x'}(\theta)$ with $\mathcal{D}'$ according to (21). When K episodes are finished, we adapt the meta-model θ_{meta} to the final model θ^* by the following function

$$\theta^* \leftarrow \theta - \beta\nabla_\theta \sum\nolimits_{\mathcal{T}_x'} L_{\mathcal{T}_x'}(\theta_{\text{meta}}) \quad (23)$$

where β is the meta-adaptation step size.

Algorithm 1. Meta mean field Q-learning.

Input: the learning rate α, the meta-adaptation step size β.
Output: the channel selection policy $\boldsymbol{\pi}_*^n$.
Initialize: the model parameters θ
Training phase:
Sample tasks as $\mathcal{T} \sim p(\tau)$.
for $k = 0 \rightarrow (K-1)$ **do**
Sample a task $\mathcal{T}_x \in \mathcal{T}$ and reset the UAV communication network.
for $t = 0 \rightarrow (E-1)$ **do**
Each UAV transmitter obtains $\mathbf{o}_t^n$ and $\bar{a}^n$, and selects the communication channel a_t^n.
Each UAV communicates over its selected channel and receives the reward r_t and a new observation $\mathbf{o}_{t+1}$.
Each UAV transmitter calculates a new average action $\bar{a}^n$ and stores the trajectory $\langle \mathbf{o}_t^n, a_t^n, r_t^n, \mathbf{o}_{t+1}^n, \bar{a}_t^n \rangle$ to the experience pool.
if Experience pool is full **then**
Draw a small batch of samples $\langle \mathbf{o}_t^n, a_t^n, r_t^n, \mathbf{o}_{t+1}^n, \bar{a}^n \rangle$ from the experience pool denoted as $\mathcal{D}$.
Evaluate $\nabla_\theta L_{\mathcal{T}_x}(\theta)$ using $\mathcal{D}$ and (21).
Update $\theta_{\text{meta}} \leftarrow \theta - \alpha \nabla_\theta L_{\mathcal{T}_x}(\theta)$.
end if
end for
end for
Return θ_{meta}
Adaptation phase:
Sample tasks as $\mathcal{T}_{\text{new}} \sim p(\tau)$.
for $k = 0 \rightarrow (K-1)$ **do**
Sample a task $\mathcal{T}_x' \in \mathcal{T}_{\text{new}}$.
for $t = 0 \rightarrow (E-1)$ **do**
Repeat step 6)–8) using θ_{meta}.
if Experience pool is full **then**
Draw a small batch of samples $\langle \mathbf{o}_t^n, a_t^n, r_t^n, \mathbf{o}_{t+1}^n, \bar{a}_t^n \rangle$ from the experience pool denoted as $\mathcal{D}'$.
Evaluate $\nabla_\theta L_{\mathcal{T}_x'}(\theta)$ using $\mathcal{D}'$and (21).
end if
end for
end for
Update $\theta^* \leftarrow \theta - \beta \nabla_\theta \sum_{\mathcal{T}_x'} L_{\mathcal{T}_x'}(\theta_{\text{meta}})$.
Return θ^*

6. Theoretical Analysis of MMFQ

6.1 Convergence Analysis

The Q-function of each UAV pair is visited infinitely open for update, and the reward is bounded by a constant. In addition, the UAV pair's policy is Greedy with the Boltzmann policy, where the Q-function is in the limit as the temperature decays to zero. Therefore, we can prove that the proposed algorithm converges in the training phase, i.e., the joint policy $\boldsymbol{\pi}_t \triangleq \left[\pi_t^1, \ldots, \pi_t^n\right]$ converges to the Nash equilibrium $\boldsymbol{\pi}_t^* \triangleq \left[(\pi_t^1)^*, \ldots, (\pi_t^n)^*\right]$ within the finite update episodes [24].

Next, we prove the loss function is bounded in the meta-adaptation phase. The parameters of model θ are adapted across new jamming policies $\mathcal{T}_x' \in \mathcal{T}_{\text{new}}$, and the objective of the meta-adaptation phase can be described as

$$\min \sum_{\mathcal{T}_x'} L_{\mathcal{T}_x'}(\theta_{\text{meta}}). \tag{24}$$

For the sake of illustration, we denote $\sum_{\mathcal{T}_x'} L_{\mathcal{T}_x'}(\theta_{\text{meta}})$ as $L(\theta_{\text{meta}})$ and make the following assumptions about the loss function $L(\theta)$ [26]. Assume the loss function $L(\theta)$ is μ-strongly convex and H-smooth, which is:

Assumption 1 *The loss function $L(\theta)$ is μ-strongly convex and H-smooth, and there is a constant B satisfied with $\|\nabla L(\theta)\| \leq B$. The Hessian of $L(\theta)$ is ρ-Lipschitz.*

Assumption 2 *There exists constants δ and σ such that*

$$\left\|\nabla L_x(\theta) - \nabla L_y(\theta)\right\| \leq \delta, \left\|\nabla^2 L_x(\theta) - \nabla^2 L_y(\theta)\right\| \leq \sigma.$$

Assumption 1 is typical in machine learning, which guarantees that the loss function $L(\theta)$ is smooth and bounded. Moreover, the Hessian smoothness of the loss function enables the characterization of the meta-adaptation objective function. Assumption 2 indicates the similarity between different tasks. δ and σ can be adjusted over a large range, and smaller constants show higher task similarity, and vice versa.

To prove the convergence of MMFQ, we first analyze the boundary properties of the loss function $L(\theta)$. Then, we discuss the influence of task similarity on convergence.

Lemma 1 *Suppose that the assumptions above hold. When $\alpha \leq \min\left(\frac{\mu}{2\mu H + \rho B}, \frac{1}{\mu}\right)$, $L(\theta_{\text{meta}})$ is p-strongly convex and q-smooth, where $p = \mu(1-\alpha H)^2 - \alpha\rho B > 0$ and $q = H(1-\alpha H)^2 + \alpha\rho B$.*

Proof 1 *See Appendix A.*

Lemma 1 demonstrates the objective function of meta-adaptation $L(\theta_{\text{meta}})$ is smooth and bounded as well as the loss function $L(\theta)$ when the learning rate α is satisfied with the value condition.

Theorem 1 *Suppose that the assumptions above hold. There exists the following expression.*

$$\left\|\nabla L_{\mathcal{T}_x}(\theta_{\text{meta}}) - \nabla L(\theta_{\text{meta}})\right\| \leq \delta + \alpha C(H\delta + B\sigma + \tau)$$

where C is a constant, and $\tau = \delta\sigma$.

Proof 2 *See Appendix B.*

Theorem 1 establishes that with the assumption of bounded variance of the gradient and Hessian of $L(\theta)$, it is possible to derive an upper bound on the objective function for meta-adaptation when faced with task variability.

According to the illustration above, the new theorem about the convergence of MMFQ can be obtained as follows.

Theorem 2 *Suppose that Assumption 1 and Assumption 2 hold, and the learning rate α and adapting step size β are satisfied with $\alpha \leq \min\left(\frac{\mu}{2\mu H+\rho B}, \frac{1}{\mu}\right)$ and $\beta < \min\left\{\frac{1}{2p}, \frac{2}{q}\right\}$. We have*

$$L\left(\theta^{E}\right) - L\left(\theta^{*}\right) \leq \xi^{nE}\left[L\left(\theta^{0}\right) - L\left(\theta^{*}\right)\right] + \frac{(1-\alpha\mu)B}{1-\xi^{E_0}} h(E_0)$$

where $\xi = 1 - 2\beta p(1 - \frac{\beta q}{2})$, $h(x) \triangleq \frac{\alpha'}{\beta q}(1+\beta q)^{t-(n-1)E} - \alpha'[t-(n-1)E]$, and $\alpha' = \beta[\delta + \alpha C(H\delta + B\sigma + \tau)]$.

Proof 3 *See Appendix C.*

From Theorem 2, we observe that $h(x)$ increases with δ, which represents the convergence impact of task similarity and update steps E_0, i.e., for a given time E, higher task similarity and smaller update steps leading to a lower convergence error.

6.2 Fast Adaptation Performance Evaluation

The adaptation performance of MMFQ for different jamming environments depends on the size of the sample data and the task similarity. The distance between the output model of reinforcement learning θ_{meta} and the optimal model of meta-learning θ^* is assumed to be bounded by ζ, which represents the convergence error. Furthermore, we define the average loss over different sample tasks in the adaptation phase as $L_\tau(\theta)$ over the task distribution p as

$$L_\tau(\theta) \triangleq \mathbb{E}_{\tau\sim p} l(\theta, \mathcal{T}_x). \tag{25}$$

Denote $L_a(\theta)$ as the sample average approximation of $L_\tau(\theta)$. Define $\chi = \theta_{\text{meta}} - \alpha\nabla L_a(\theta_{\text{meta}})$ and $\chi^* = \arg\min L_{\mathcal{T}}(\theta) = \theta_a^* - \alpha\nabla L_\tau(\theta_a^*)$. Then we can derive the following theorem that characterizes the impact of task similarity on adaptive performance.

Theorem 3 *For any $\zeta > 0$, there exists a positive constant C_a and $n = n(\zeta)$ satisfying the following expression with probability at least $\left(1 - C_a e^{-Kn}\right)$,*

$$\begin{aligned}\|L_\tau(\chi) - L_\tau(\chi^*)\| \leq{}& \alpha H\zeta + H(1+\alpha H)\zeta \\ &+ H(1+\alpha H)\|\theta^* - {\theta_a}^*\|.\end{aligned}$$

Proof 4 *See Appendix D.*

The difference in performance between the output of the training phase and the optimal model is upper bounded by $\|\theta^* - {\theta_a}^*\|$, which demonstrates that higher task similarity leads to better performance. The difference among tasks, i.e., the jammers' location, exactly satisfies the higher task similarity.

7. Experimental Results

In this section, we first comprehensively and comparatively evaluate our proposed algorithm against several state-of-the-art algorithms, i.e., Mean-Field Q-learning (MFQ) [19], Probabilistic Q-learning (PQ) [27], and Independent Q-Learning (IQL) [28], all with the same simulation design. MFQ lacks an adaptation phase compared to the proposed algorithm and requires complete retraining of the model when the jamming environment changes. PQ and IQL are different Q-learning without mean-filed theory. Specifically, PQ allows the UAV to choose a channel based on probability distributions instead of the maximum Q-value, whereas IQL learns and updates regardless of the influence of other UAVs' behavior on its decisions. We then verify the effect of variations in the number of training and adaptation tasks on performance. The simulation results demonstrate the MMFQ converges faster and achieves higher throughput and faster adaptation to new tasks.

Experimental settings: Consider the initial environment with 100 UAV pairs, i.e., 200 UAVs, and 2 jammers. The number of available channels is 80. In the adaptation phase, θ_{meta} are tested for sample tasks after 200 episodes, and each episode contains 2000 steps. In addition, MMFQ contains a training phase with episodes and steps being parallel to the adaptation phase. All the experiments were executed on Pycharm with Python 3.8, and Torch 1.6 was adopted as the neural network framework. To demonstrate the convergence of the algorithm in a visible way, the learning rate α and the adaptation rate β are both set as 0.01. The discount factor $\gamma = 0.95$. The other experimental settings are listed in Tab. 1, which were set according to commonly actual scenarios [29–31].

Convergence performance: Figure 4 plots the experimental results of the mean reward for different numbers of episodes. First, we observe that the proposed algorithm converges before the 40th episode, attaining a final mean reward value of approximately 0.9. Although the MFQ converges at approximately the 40th episode, the MMFQ has a significant increase at the 10th episode and is smoother after convergence. Therefore, MMFQ can converge faster and yield a superior mean reward with the adaptation phase when it comes to new tasks. Second, it can be seen that MMFQ outperforms MFQ, MFQ outperforms PQ, and PQ outperforms IQL, indicating that MMFQ takes advantage of both mean-field and meta-learning, enhancing its capacity to select the spectrum more effectively.

Parameter	Value
Transmit power of UAV (P_{i})	23 dBm
Transmit time of one time slot (T_{trans})	0.98 s
Channel bandwidth (B)	1.5 MHz
Jamming power (P_{j})	23 dBm
Noise power (N_{n})	–114 dBm
The maximum distance within a UAV pair ($d_{n,n}$)	50 m
The threshold distance of neighbors (d_{th})	100 m

Tab. 1. Parameter settings.

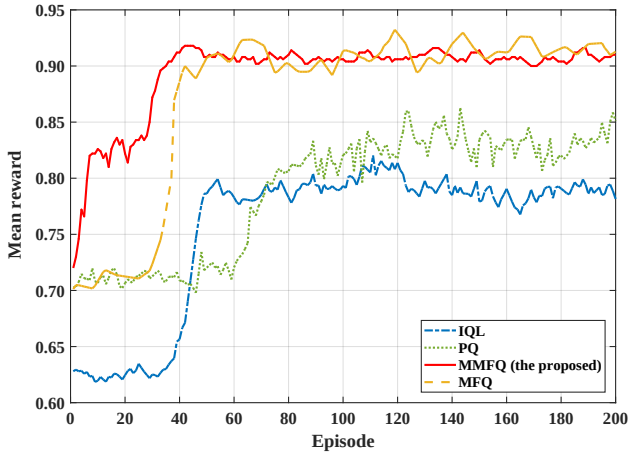

Fig. 4. The convergence performance of different algorithms for the new task.

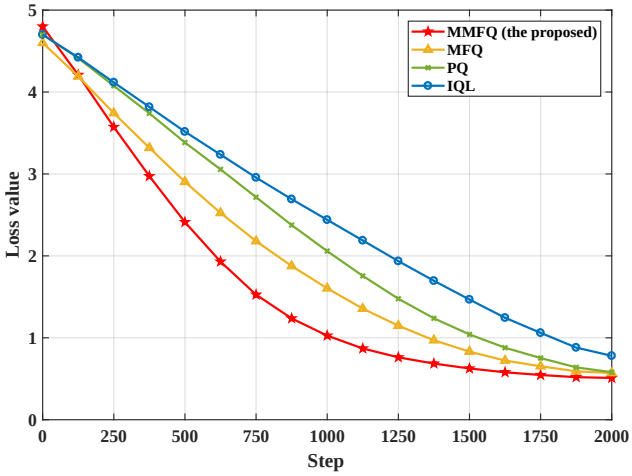

Fig. 5. The loss value comparison of different algorithms varies with the steps for each episode.

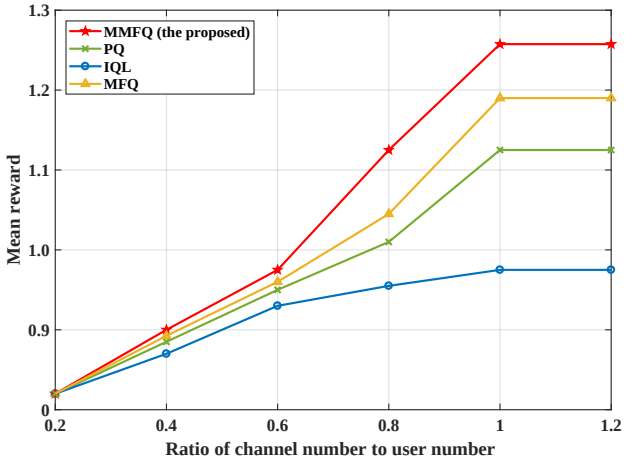

Fig. 6. The mean reward of different algorithms varies with the number of channels.

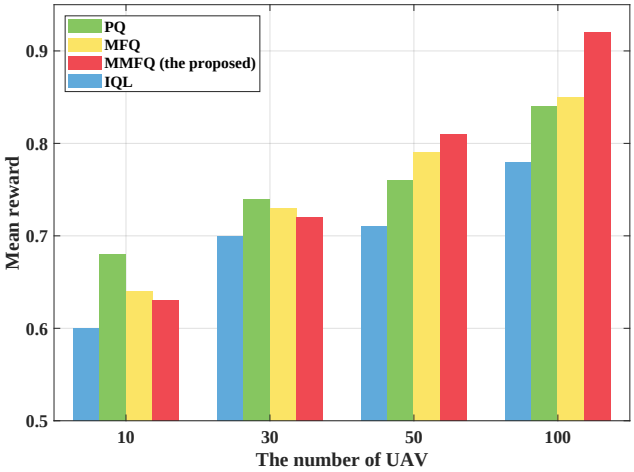

Fig. 7. The mean reward of different algorithms varies with the number of UAV.

Performance in each episode: Figure 5 illustrates the variation trend of loss values in the communication system during each episode with 2000 steps. First, we can figure out that the slope of the MMFQ trajectory achieves maximum value earlier than other algorithms, and the decline trajectories of other algorithms are smoother. Second, we find that MMFQ can reach a lower final loss value at each episode, enabling the MMFQ to converge faster and ultimately gain a higher average reward than PQ and IQL. Since MMFQ enables the meta-model to yield significant improvement with minor adjustments, accelerating the process of convergence under a new jamming environment.

With different spectrum resources: Figure 6 shows the obtained mean reward with different spectrum resources, where the horizontal coordinate represents the ratio of channels to UAVs, i.e., an increase in the number of channels. The number of UAVs is fixed at 200, and the number of available channels increases gradually from 40 to 240. We can notice that where the quantity of available channels does not comprehensively match the number of UAVs, the growth tendency of the mean reward achieved by MMFQ lies between linear and exponential growth until the number of channels exceeds the number of UAVs. In contrast, the mean reward for the benchmark algorithms is growing slowly, which indicates that the proposed algorithm offers UAVs an enhanced capability to effectively coordinate spectrum resources while mitigating the risk of co-channel interference.

With different number of users: Figure 7 plots the experimental results of the mean reward for different numbers of UAVs (10, 30, 50, and 100). The number of channels is accordingly set to one-half of the number of UAVs. We can observe that the advantage of the proposed algorithm becomes more apparent as the number of users increases. When the number of users is 10 and 20, PQ and MFQ perform slightly better than MMFQ. While the number of users rises to 100, MMFQ significantly outperforms the benchmark algorithms in terms of mean reward. The experimental results demonstrate that MMFQ can achieve better performance in large-scale communication systems.

Throughput analysis: Figure 8 shows the regression of the throughput achieved by the system under various algorithms. First, we can observe that the slope of the regression curve for MMFQ, MFQ, PQ, and IQL are 0.175, 0.1, 0.065, and 0.035 respectively, appearing as a greater boost in throughput and a higher ultimate value by utilizing MMFQ. In contrast, the benchmark algorithms have a relatively smooth and slow increase in throughput. Obviously, with the assistance of training tasks, the agent can leverage the training result to adjust the meta-model more effectively. Second, we notice that the data points of MMFQ are more discrete than the benchmark algorithms, indicating the throughput is unstable in the initial part of the adaptation phase. The larger fluctuations, however, ultimately higher values, demonstrate that MMFQ can leverage the model to adapt to new tasks more effectively.

Adaptation performance: The experiments conducted before all with the same task setting for MMFQ, i.e., 30 training tasks and 30 adaptation tasks. MFQ, IQL, and PQ are without training tasks and directly processing with new tasks. To further evaluate the adaptive performance of MMFQ, we conduct different task settings for MMFQ as shown in Tab. 2. Note that the IQL and PQ have similar comparison effects for MMFQ, therefore we compare the adaptation performance with MFQ and PQ under the conditions that the number of UAVs $N = 50$ and channels $M = 40$ as shown in Fig. 9. First, we observe that MMFQ and MFQ outperform PQ across all task settings even if the scale of training or adaptation tasks is small. Second, we find that the difference between MMFQ and MFQ is insignificant in Set A, which indicates that the UAV cannot effectively utilize the meta-model when the number of training tasks is small. With the increasing number of training tasks, MMFQ converges faster than MFQ in Set B, showing that the meta-model has been well-trained in the training phase. In Set C, MMFQ significantly outperforms the MFQ, including faster convergence and more stable trends. The reason is that MMFQ can quickly update the meta-model to the optimal one with minor adjustments when performing too many adaptation tasks, enabling the UAVs' enhanced adaptability toward new jamming environments.

Type	Training Tasks	Adaptation Task
Set A	2	2
Set B	30	2
Set C	30	30

Tab. 2. Different task settings for MMFQ.

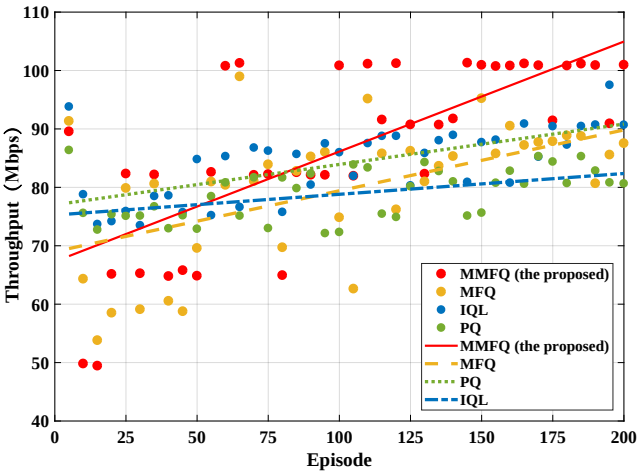

Fig. 8. The throughput comparison of different algorithms varies with the steps.

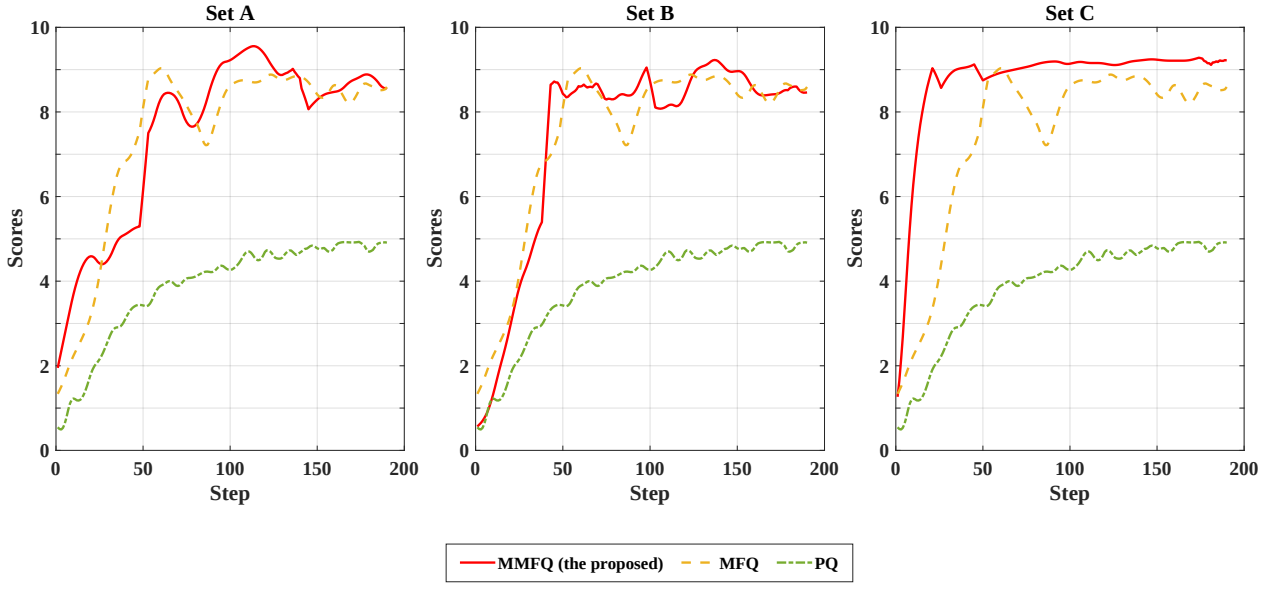

Fig. 9. Meta-adaptation performance comparison of MMFQ with different training and adaptation tasks.

8. Conclusion

In this paper, we have proposed an anti-jamming channel selection scheme for a large-scale UAV communication network, where each UAV pair aims to maximize its long-term expected achievable throughput by finding the optimal channel selection policy with malicious jammers and co-channel interference. Furthermore, we have designed a POSG to formulate the channel selection process and utilized the mean-field theory to simplify the computational complexity. Then, we have developed an algorithm based on MMFQ to cultivate adaptability and promote self-oriented exploration when facing novel tasks. Moreover, we have proved the convergence and the fast adaptation performance of the proposed algorithm. Experimental results show that the proposed algorithm converges faster, achieves higher throughput, and adapts to new tasks more quickly than the benchmark algorithms, especially for large-scale communication networks. Exploring more appropriate and effective methods for incorporating the MMFQ within the broader framework of the UAV communication system represents an interesting avenue for future research.

Acknowledgments

The mathematical model can be available at: https://github.com/H953242/Meta-RL-for-multi-UAV.git

References

[1] LIU, D., XU, Y., WANG, J., et al. Opportunistic UAV utilization in wireless networks: Motivations, applications, and challenges. *IEEE Communications Magazine*, 2020, vol. 58, no. 5, p. 62–68. DOI: 10.1109/MCOM.001.1900687

[2] WANG, Z., DUAN, L. Chase or wait: Dynamic UAV deployment to learn and catch time-varying user activities *IEEE Transactions on Mobile Computing*, 2021, vol. 22, no. 3, p. 1369–1383. DOI: 10.1109/TMC.2021.3107027

[3] LIU, Q., SHI, L., SUN, L., et al. Path planning for UAV-mounted mobile edge computing with deep reinforcement learning. *IEEE Transactions on Vehicular Technology*, 2020, vol. 69, no. 5, p. 5723–5728. DOI: 10.1109/TVT.2020.2982508

[4] VISHWAKARMA, N. K., SINGH, R. K. Design and implementation of FHSS (Frequency Hopping Spread Spectrum) synthesizer. In *7th International Conference on Signal Processing and Communication (ICSC)*. Noida (India), 2021, p. 151–155. DOI: 10.1109/ICSC53193.2021.9673302

[5] RAJARAJESWARIE, B., SANDANALAKSHMI, R. An adaptive beamforming algorithm based on FPGA synthesis for MIMO antennas. In *International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)*. Bangalore (India), 2022, p. 1–4. DOI: 10.1109/SMARTGENCON56628.2022.10084306

[6] JIA, K., CHEN, D., SUN, X. Soft actor-critic based power control algorithm for anti-jamming in D2D communication. In *IEEE International Conference on Control, Electronics and Computer Technology (ICCECT)*. Jinlin (China), 2023, p. 1–5. DOI: 10.1109/ICCECT57938.2023.10140869

[7] WEI, F., ZHENG, S., ZHOU, X., et al. Detection of direct sequence spread spectrum signals based on deep learning. *IEEE Transactions on Cognitive Communications and Networking*, 2022, vol. 8, no. 3, p. 1399–1410. DOI: 10.1109/TCCN.2022.3174609

[8] LIU, S., XU, Y., CHEN, X., et al. Pattern-aware intelligent anti-jamming communication: A sequential deep reinforcement learning approach. *IEEE Access*, 2019, vol. 7, p. 169204–169216. DOI: 10.1109/ACCESS.2019.2954531

[9] CHANG, X., LI, Y., ZHAO, Y., et al. An improved anti-jamming method based on deep reinforcement learning and feature engineering. *IEEE Access*, 2022, vol. 10, p. 69992–70000. DOI: 10.1109/ACCESS.2022.3187030

[10] YIN, Z., LIN, Y., ZHANG, Y., et al. Collaborative multiagent reinforcement learning aided resource allocation for UAV anti-jamming communication. *IEEE Internet of Things Journal*, 2022, vol. 9, no. 23, p. 23995–24008. DOI: 10.1109/JIOT.2022.3188833

[11] LI, Z., LU, Y., LI, X., et al. UAV networks against multiple maneuvering smart jamming with knowledge-based reinforcement learning. *IEEE Internet of Things Journal*, 2021, vol. 8, no. 15, p. 12289–12310. DOI: 10.1109/JIOT.2021.3062659

[12] ZHANG, Y., JIA, L., QI, N., et al. A multi-agent reinforcement learning anti-jamming method with partially overlapping channels. *IET Communications*, 2021, vol. 15, no. 19, p. 2461–2468. DOI: 10.1049/CMU2.12288

[13] XU, H., WU, J., PAN, Q., et al. Digital twin and meta RL empowered fast-adaptation of joint user scheduling and task offloading for mobile industrial IoT. *IEEE Journal on Selected Areas in Communications*, 2023, vol. 41, no. 10, p. 3254–3266. DOI: 10.1109/JSAC.2023.3310081

[14] YUAN, Y., ZHENG, G., WONG, K. K., et al. Meta-reinforcement learning based resource allocation for dynamic V2X communications. *IEEE Transactions on Vehicular Technology*, 2021, vol. 70, no. 9, p. 8964–8977. DOI: 10.1109/TVT.2021.3098854

[15] ZHANG, Z., WANG, N., WU, H., et al. MR-DRO: A fast and efficient task offloading algorithm in heterogeneous edge/cloud computing environments. *IEEE Internet of Things Journal*, 2021, vol. 10, no. 4, p. 3165–3178. DOI: 10.1109/JIOT.2021.3126101

[16] WAN, J., LIN, S., ZHANG, Z., et al. Scheduling real-time wireless traffic: A network-aided offline reinforcement learning approach. *IEEE Internet of Things Journal*, 2023, vol. 10, no. 24, p. 22331–22340. DOI: 10.1109/JIOT.2023.3304969

[17] FERIANI, A., WU, D., XU, Y., et al. Multiobjective load balancing for multiband downlink cellular networks: A meta-reinforcement learning approach. *IEEE Journal on Selected Areas in Communications*, 2022, vol. 40, no. 9, p. 2614–2629. DOI: 10.1109/JSAC.2022.3191114

[18] HUANG, M., CAINES, P. E., MALHAME, R. P. The NCE (mean field) principle with locality dependent cost interactions. *IEEE Transactions on Automatic Control*, 2010, vol. 55, no. 12, p. 2799–2805. DOI: 10.1109/TAC.2010.2069410

[19] SUN, Y., LI, L., CHENG, Q., et al. Joint trajectory and power optimization in multi-type UAVs network with mean field Q-learning. In *IEEE International Conference on Communication Workshops (ICC Workshops)*. Dublin (Ireland), 2020, p. 1–6. DOI: 10.1109/ICCWorkshops49005.2020.9145105

[20] SUN, Y., LI, L., XUE, K., et al. Inhomogeneous multi-UAV aerial base stations deployment: A mean-field-type game approach. In *15th International Wireless Communication and Mobile Computing Conference (IWCMC)*. Tangier (Morocco), 2019, p. 1204–1208. DOI: 10.1109/IWCMC.2019.8766540

[21] SHIRI, H., PARK, J., BENNIS, M. Massive autonomous UAV path planning: A neural network based mean-field game theoretic approach. In *IEEE Global Communication Conference (GLOBECOM)*. Waikoloa (HI, USA), 2019, p. 1–6. DOI: 10.1109/GLOBECOM38437.2019.9013181

[22] WANG, X., XU, Y., CHEN, J., et al. Mean field reinforcement learning based anti-jamming communications for ultra-dense internet of things in 6G. In *International Conference on Wireless Communications and Signal Processing (WCSP)*. Nanjing (China), 2020, p. 195–200. DOI: 10.1109/WCSP49889.2020.9299742

[23] LI, D., ZHOU, J., WANG, J., et al. Linking generation rate based on Gauss-Markov mobility model for mobile ad hoc networks. In *International Conference on Networks Security, Wireless Communications and Trusted Computing*. Wuhan (China), 2009, p. 358–361. DOI: 10.1109/NSWCTC.2009.286

[24] YANG, Y., LUO, R., LI, M., et al. Mean field multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*. Stockholmsmassan (Sweden), 2018, p. 5571–5580. DOI: 10.48550/arXiv.1802.05438

[25] FINN, C., ABBEEl, P., LEVINE, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*. Sydney (Australia), 2017, p. 1126–1135. DOI: 10.5555/3305381.3305498

[26] LIN, S., YANG, G., ZHANG, J. Real-time edge intelligence in the making: A collaborative learning framework via federated meta-learning. *arXiv*, 2020, p. 1–13. DOI: 10.48550/arXiv.2001.03229

[27] CHEN, C., DONG, D., LI, H., et al. Fidelity-based probabilistic Q-learning for control of quantum systems. *IEEE Transactions on Neural Networks and Learning Systems*, 2014, vol. 25, no. 5, p. 920–933. DOI: 10.1109/TNNLS.2013.2283574

[28] ZHOU, Y., ZHOU, F., WU, Y., et al. Subcarrier assignment schemes based on Q-learning in wideband cognitive radio networks. *IEEE Transactions on Vehicular Technology*, 2020, vol. 69, no. 1, p. 1168–1172. DOI: 10.1109/TVT.2019.2953809

[29] NADEEM, A., ULLAH, A., CHOI, W. Social-aware peer selection for energy efficient D2D communications in UAV-assisted networks: A Q-learning approach. *IEEE Wireless Communications Letters*, 2024, vol. 13, no. 5, p. 1468–1472. DOI: 10.1109/LWC.2024.3375235

[30] WANG, J., MA, Y., LU, R., et al. Hovering UAV-based FSO communications: Channel modeling, performance analysis, and parameter optimization. *IEEE Journal on Selected Areas in Communications*, 2021, vol. 39, no. 10, p. 2496–2959. DOI: 10.1109/JSAC.2021.3088656

[31] LI, A., ZHANG, W. Mobile jammer-aided secure UAV communications via trajectory design and power control. *China Communications*, 2018, vol. 15, no. 8, p. 141–151. DOI: 10.1109/CC.2018.8438280

About the Authors ...

Linzi HU received the B.S. degree from the School of Electronic and Optical Engineering, at Nanjing University of Science and Technology in 2021, where she is currently working toward the M.S. degree. Her research interests include reinforcement learning, deep learning, and UAV communication.

Yumeng SHAO received the B.S. degree from the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China, in 2019, where he is currently working toward the Ph.D. degree. His research interests include distributed machine learning, blockchain, game theory, and trusted AI.

Yuwen QIAN (corresponding author) was born in 1975. He received his Ph.D. degree in Automatic Engineering from Nanjing University of Science and Technology, Nanjing, China, in 2011. From 2002 to 2011, he was a Lecturer at the School of Automation, Nanjing University of Science and Technology. Since 2019, he has been an Associate Professor with the School of Electronic and Optical Engineering, at Nanjing University of Science and Technology.

Feng DU received the B.S. and M.S. degree from the School of Electronic and Optical Engineering, at Nanjing University of Science and Technology in 2020 and 2023, respectively. His research interests include reinforcement learning and UAV communication.

Jun LI received a Ph.D. degree in Electronic Engineering from Shanghai Jiao Tong University, Shanghai, China, in 2009. From January 2009 to June 2009, he was a Research Scientist at the Department of Research and Innovation, Alcatel Lucent Shanghai Bell, Shanghai. From June 2009 to April 2012, he was a Post-Doctoral Fellow at the School of Electrical Engineering and Telecommunications, University of New South Wales, Kensington, Australia. From April 2012 to June 2015, he was a Research Fellow at the School of Electrical Engineering, University of Sydney, Camperdown, Australia. He was a Visiting Professor at Princeton University, Princeton, NJ, USA, from 2018 to 2019. Since 2015, he has been a Professor at the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China.

Yan LIN received the M.S. and Ph.D. degrees from the Southeast University, China, in 2013 and 2018, respectively. She visited the Southampton Wireless Group, Southampton University, U.K., from October 2016 to October 2017. She joined Nanjing University of Science and Technology, China, in 2018, where she is currently an Associate Professor with the School of Electronic and Optical Engineering. Her current research interests include vehicular networks, UAV networks, mobile edge computing, and reinforcement learning for resource allocation in wireless communication.

Zhe WANG received the B.S. degree in Electrical Engineering and Automation and the M.S. degree in Control Science and Engineering from the North China University of Technology, Beijing, China, in 2015 and 2018, respectively. She is currently pursuing the Ph.D. degree in Electrical Engineering with the University of Louisville, Louisville, KY, USA. Since 2018, she has been a Graduate Research Assistant with the Department of Electrical and Computer Engineering, University of Louisville. Her current research interests include wireless communication, aeronautical ad hoc networks, deep learning, and resource allocation. She was a recipient of the Best of Track Award from IEEE/AIAA 40th Digital Avionics Systems Conference (DASC), in 2021.

Appendix A: Proof of Lemma 1

We demonstrate the objective of meta-adaptation $L(\theta_{\text{meta}})$ is p-strongly convex and q-smooth. According to (22), we have

$$\nabla_\theta L(\theta_{\text{meta}}) = \nabla_\theta L(\theta_{\text{meta}}) - \alpha \nabla_\theta L(\theta_{\text{meta}}) \cdot \nabla_\theta^2 L(\theta). \quad \text{(A1)}$$

Therefore, we can deduce that

$$\begin{aligned} &\|\nabla_\theta L(\theta_{\text{meta}}) - \nabla_\theta L(\theta'_{\text{meta}})\| \\ &= \left\| \left(1 - \alpha \nabla_\theta^2 L(\theta)\right) \left[\nabla_\theta L(\theta_{\text{meta}}) - \nabla_\theta L(\theta'_{\text{meta}})\right] \right. \\ &\quad \left. - \alpha \nabla_\theta L(\theta'_{\text{meta}}) \left[\nabla_\theta^2 L(\theta) - \nabla_\theta^2 L(\theta')\right] \right\|. \end{aligned} \quad \text{(A2)}$$

With the Assumptions 1, Equation (A2) can be derived as follows

$$\begin{aligned} &\|\nabla_\theta L(\theta_{\text{meta}}) - \nabla_\theta L(\theta'_{\text{meta}})\| \\ &\geq (1 - \alpha H) \|\nabla_\theta L(\theta_{\text{meta}}) - \nabla_\theta L(\theta'_{\text{meta}})\| \\ &\quad - \alpha \left\| \nabla_\theta L(\theta'_{\text{meta}}) \left[\nabla_\theta^2 L(\theta) - \nabla_\theta^2 L(\theta')\right] \right\| \\ &\geq \mu (1 - \alpha H) \|\theta_{\text{meta}} - \theta'_{\text{meta}}\| - \alpha \rho B \|\theta - \theta'\|. \end{aligned} \quad \text{(A3)}$$

From $\nabla_\theta^2 L(\theta) = \frac{\nabla_\theta L(\theta) - \nabla_\theta L(\theta')}{\|\theta - \theta'\|} \leq H$, we can obtain the expression

$$1 - \alpha H \leq 1 - \alpha \nabla_\theta^2 L(\theta) \leq 1 - \alpha \mu. \quad \text{(A4)}$$

With respect to $\nabla_\theta \theta_{\text{meta}} = \frac{\|\theta_{\text{meta}} - \theta'_{\text{meta}}\|}{\|\theta - \theta'\|}$, we have

$$(1 - \alpha H) \|\theta - \theta'\| \leq \|\theta_{\text{meta}} - \theta'_{\text{meta}}\| \leq (1 - \alpha \mu) \|\theta - \theta'\|. \quad \text{(A5)}$$

According to (A3), (A4) and (A5), (A2) can be restricted by the following expression

$$p \|\theta - \theta'\| \leq \|\nabla_\theta L(\theta_{\text{meta}}) - \nabla_\theta L(\theta'_{\text{meta}})\| \leq q \|\theta - \theta'\| \quad \text{(A6)}$$

where $p = \mu(1 - \alpha H)^2 - \alpha \rho B > 0$ and $q = H(1 - \alpha H)^2 + \alpha \rho B$.

Thus, if $\alpha \leq \min\left(\frac{\mu}{2\mu H + \rho B}, \frac{1}{\mu}\right)$, $L(\theta_{\text{meta}})$ is p-strongly convex and q-smooth, where $p = \mu(1 - \alpha H)^2 - \alpha \rho B > 0$ and $q = H(1 - \alpha H)^2 + \alpha \rho B$.

Appendix B: Proof of Theorem 1

From Taylor's theorem, it follows that

$$\begin{aligned} \nabla L_{\mathcal{T}_x}(\theta_{\text{meta}}) &= \nabla L_{\mathcal{T}_x}(\theta) + \nabla^2 L_{\mathcal{T}_x}(\theta)(\theta_{\text{meta}} - \theta) \\ &\quad + o\left(\|\theta_{\text{meta}} - \theta\|^2\right). \end{aligned} \quad \text{(B1)}$$

According to (22) and Assumption 1, Equation (B1) can be rewritten as

$$\nabla L_{\mathcal{T}_x}(\theta_{\text{meta}}) = \nabla L_{\mathcal{T}_x}(\theta) - \alpha \nabla^2 L_{\mathcal{T}_x}(\theta) \nabla L_{\mathcal{T}_x}(\theta) + o\left(\alpha^2 B^2\right). \quad \text{(B2)}$$

Based on Assumption 1 and Assumption 2, the product between the Hessian matrix and the gradient of $L_{\mathcal{T}_x}(\theta)$ can be expressed as follows

$$\left\|\nabla^2 L_{\mathcal{T}_x}(\theta) \nabla L_{\mathcal{T}_x}(\theta) - \nabla^2 L(\theta_{\text{meta}}) \nabla L(\theta_{\text{meta}})\right\| \leq \omega \quad \text{(B3)}$$

where $\omega = H\delta + B\sigma + \tau$, and $\tau = \delta\sigma$.

Similarly, we can deduce the following expression

$$\left\|\left[\nabla^2 L_{\mathcal{T}_x}(\theta)\right]^2 \nabla L_{\mathcal{T}_x}(\theta) - \left[\nabla^2 L(\theta_{\text{meta}})\right]^2 \nabla L(\theta_{\text{meta}})\right\| \leq \omega' \quad \text{(B4)}$$

where $\omega' = H\delta' + B\sigma + \tau'$, $\delta' = H\delta + B\sigma + \tau$, and $\tau' = \delta'\sigma$.

Based on (B3) and (B4), we have

$$\begin{aligned} &\left\|\nabla L_{\mathcal{T}_x}(\theta_{\text{meta}}) - \nabla L(\theta_{\text{meta}})\right\| \\ &= \left\| \left[I - \alpha \nabla^2 L_{\mathcal{T}_x}(\theta)\right] \left[\nabla L_{\mathcal{T}_x}(\theta) - \nabla L_{\mathcal{T}_j}(\theta) + \nabla L_{\mathcal{T}_j}(\theta)\right] \right. \\ &\quad \left. - \sum_{\mathcal{T}} \left[I - \alpha \nabla^2 L_{\mathcal{T}_x}(\theta)\right] \left[\nabla L_{\mathcal{T}_x}(\theta) - \nabla L_{\mathcal{T}_j}(\theta) + \nabla L_{\mathcal{T}_j}(\theta)\right] \right\| \\ &= \left\| \nabla L_{\mathcal{T}_x}(\theta) - \nabla L_{\mathcal{T}_j}(\theta) - 2\alpha \nabla^2 L_{\mathcal{T}_x}(\theta) \nabla L_{\mathcal{T}_x}(\theta) \right. \\ &\quad + 2\alpha \sum_{\mathcal{T}} \nabla^2 L_{\mathcal{T}_x}(\theta) \nabla L_{\mathcal{T}_x}(\theta) + o\left(\alpha^2 B^2\right) \\ &\quad \left. + \alpha^2 \left[\nabla^2 L_{\mathcal{T}_x}(\theta)\right]^2 \nabla L_{\mathcal{T}_x}(\theta) - \alpha^2 \sum_{\mathcal{T}} \left[\nabla^2 L_{\mathcal{T}_x}(\theta)\right]^2 \nabla L_{\mathcal{T}_x}(\theta) \right\| \\ &\leq \left\|\nabla L_{\mathcal{T}_x}(\theta) - \nabla L_{\mathcal{T}_j}(\theta)\right\| + o\left(\alpha^2 B^2\right) \\ &\quad + 2\alpha \left\|\nabla^2 L_{\mathcal{T}_x}(\theta) \nabla L_{\mathcal{T}_x}(\theta) - \sum_{\mathcal{T}} \nabla^2 L_{\mathcal{T}_x}(\theta) \nabla L_{\mathcal{T}_x}(\theta)\right\| \\ &\quad + \alpha^2 \left\| \left[\nabla^2 L_{\mathcal{T}_x}(\theta)\right]^2 \nabla L_{\mathcal{T}_x}(\theta) - \sum_{\mathcal{T}} \left[\nabla^2 L_{\mathcal{T}_x}(\theta)\right]^2 \nabla L_{\mathcal{T}_x}(\theta) \right\| \end{aligned} \quad \text{(B5)}$$

$$\leq \delta + \alpha C (H\delta + B\sigma + \tau) \quad \text{(B6)}$$

where $\tau = \delta\sigma$.

We summarize that if α is relatively small, the above expression is satisfied with a fixed C, which means an upper bound of the objective function for MMFQ can be obtained with a given bounded variance of the gradient and the Hessian of the loss function in the presence of task variability.

Appendix C: Proof of Theorem 2

We define a virtual sequence $Z_{[n]}^t$ for the model θ^t where $[n]$ denotes the interval $[(n-1)E, nE]$, and $Z_{[n]}^{(n-1)E} = \theta^{(n-1)E}$. Then we have

$$Z_{[n]}^{t+1} = Z_{[n]}^t - \beta \nabla L\left(Z_{[n]}^t\right). \quad \text{(C1)}$$

Therefore, the distance between $Z_{[n]}^{t+1}$ and θ^{t+1} can be expressed as

$$\left\|\theta^{t+1}-Z_{[n]}^{t+1}\right\|=\left\|\theta^{t}-\beta\nabla L(\theta^{t})-Z_{[n]}^{t}+\beta\nabla L(Z_{[n]}^{t})\right\|. \tag{C2}$$

According to (A6), (C2) can be deduced as

$$\left\|\theta^{t+1}-Z_{[n]}^{t+1}\right\|\leq\left\|\theta^{t}-Z_{[n]}^{t}\right\|+\beta q\left\|Z_{[n]}^{t}-\theta^{t}\right\|. \tag{C3}$$

By induction, we can derive the following expression

$$\left\|\theta^{t}-Z_{[n]}^{t}\right\|\leq g(t-(n-1)E) \tag{C4}$$

where $g(x)\triangleq\frac{\delta+\alpha C(H\delta+B\sigma+\tau)}{q}[(1+\beta q)^{x}-1]$, and C, δ, σ, and $\tau=\delta\sigma$ are all constants.

Therefore, Equation (C2) can be further derived as

$$\left\|\theta^{t+1}-Z_{[n]}^{t+1}\right\|\leq\left\|\theta^{t}-Z_{[n]}^{t}\right\|+\alpha'\left[(1+\beta q)^{t-(n-1)E}-1\right] \tag{C5}$$

where $\alpha'=\beta[\delta+\alpha C(H\delta+B\sigma+\tau)]$.

When $t\in[(n-1)E,nE]$, we have

$$\begin{aligned}\left\|\theta^{t}-Z_{[n]}^{t}\right\|&\leq\sum\nolimits_{j=1}^{t-(n-1)E}\left\{\alpha'\left[(1+\beta q)^{t-(n-1)E}-1\right]\right\}\\&=\frac{\alpha'}{\beta q}(1+\beta q)^{t-(n-1)E}-\alpha'[t-(n-1)E]\\&\triangleq h(t-(n-1)E).\end{aligned} \tag{C6}$$

Next, we discuss the convergence of $Z_{[n]}^{t}$. Based on the p-strongly convex of $L(\theta_{\text{meta}})$, we can obtain

$$L\left(Z_{[n]}^{t+1}\right)-L\left(Z_{[n]}^{t}\right)\leq-\beta\left(1-\frac{\beta q}{2}\right)\left\|\nabla L\left(Z_{[n]}^{t}\right)\right\|^{2}. \tag{C7}$$

Similarly, based on the q-smooth of $L(\theta_{\text{meta}})$, we have

$$L\left(Z_{[n]}^{t}\right)\leq L(\theta^{*})+\frac{1}{2p}\left\|\nabla L\left(Z_{[n]}^{t}\right)\right\|^{2}. \tag{C8}$$

According to (C7) and (C8), the expression can be deduced as follows

$$L\left(Z_{[n]}^{t+1}\right)-L\left(Z_{[n]}^{t}\right)\leq-2\beta q\left(1-\frac{\beta q}{2}\right)\left[L\left(Z_{[n]}^{t}\right)-L(\theta^{*})\right] \tag{C9}$$

which can be rewritten as

$$L\left(Z_{[n]}^{t+1}\right)-L(\theta^{*})\leq\xi\left[L\left(Z_{[n]}^{t}\right)-L(\theta^{*})\right] \tag{C10}$$

and

$$\xi\left[L\left(Z_{[n]}^{t}\right)-L(\theta^{*})\right]\triangleq\left[1-2\beta q\left(1-\frac{\beta q}{2}\right)\right]\left[L\left(Z_{[n]}^{t}\right)-L(\theta^{*})\right] \tag{C11}$$

where $\xi\in(0,1)$, and $\beta<\min\left\{\frac{1}{2p},\frac{2}{q}\right\}$.

Iterations can be deduced as follows

$$\begin{aligned}L\left(Z_{[n]}^{nE}\right)-L(\theta^{*})&\leq\xi\left[L\left(Z_{[n]}^{nE-1}\right)-L(\theta^{*})\right]\\&\leq\xi^{2}\left[L\left(Z_{[n]}^{nE-2}\right)-L(\theta^{*})\right]\\&\quad\cdots\\&\leq\xi^{E}\left[L\left(Z_{[n]}^{(n-1)E}\right)-L(\theta^{*})\right]\\&=\xi^{E}\left[L\left(Z_{[n-1]}^{(n-1)E}\right)-L(\theta^{*})\right]\\&\quad+\xi^{E}\left[L\left(Z_{[n]}^{(n-1)E}\right)-L\left(Z_{[n-1]}^{(n-1)E}\right)\right].\end{aligned} \tag{C12}$$

Combining (A1) and Assumption 1–Assumption 2, we have

$$\|\nabla_{\theta}L(\theta_{\text{meta}})\|\leq(1-\alpha\mu)B. \tag{C13}$$

According to the mean value theorem $\|L(\theta)-L(\theta')\|\leq(1-\alpha\mu)B\|\theta-\theta'\|$ and (C6), we can derive

$$\begin{aligned}&L\left(Z_{[n]}^{(n-1)E}\right)-L\left(Z_{[n-1]}^{(n-1)E}\right)\\&=L(\theta^{(n-1)E})-L(Z_{[n-1]}^{(n-1)E})\\&\leq(1-\alpha\mu)B\left\|\theta^{(n-1)E}-Z_{[n-1]}^{(n-1)E}\right\|\\&\leq(1-\alpha\mu)Bh(E).\end{aligned} \tag{C14}$$

Iterations can be expressed as follows

$$\begin{aligned}&L\left(Z_{[n]}^{nE}\right)-L(\theta^{*})\\&\leq\xi^{E}\left[L\left(Z_{[n-1]}^{(n-1)E}\right)-L(\theta^{*})\right]+\xi^{E}(1-\alpha\mu)Bh(E)\\&\leq\xi^{2E}\left[L\left(Z_{[n-2]}^{(n-2)E}\right)-L(\theta^{*})\right]+(\xi^{E}+\xi^{2E})(1-\alpha\mu)Bh(E)\\&\quad\cdots\\&\leq\xi^{nE}\left[L\left(Z_{[0]}^{0}\right)-L(\theta^{*})\right]+\sum\nolimits_{j=1}^{N}\xi^{jE}(1-\alpha\mu)Bh(E)\\&=\xi^{nE}\left[L\left(\theta^{0}\right)-L(\theta^{*})\right]+\frac{(1-\alpha\mu)B}{1-\xi^{E}}h(E)\end{aligned} \tag{C15}$$

where $\xi=1-2\beta p(1-\frac{\beta q}{2})$.

In conclusion, when the learning rate α and adapting step size β are satisfied with $\alpha\leq\min\left(\frac{\mu}{2\mu H+\rho B},\frac{1}{\mu}\right)$ and $\beta<\min\left\{\frac{1}{2p},\frac{2}{q}\right\}$ respectively, $L(\theta^{E})-L(\theta^{*})\leq\xi^{nE}\left[L(\theta^{0})-L(\theta^{*})\right]+\frac{(1-\alpha\mu)B}{1-\xi^{E_0}}h(E_0)$.

Appendix D: Proof of Theorem 3

Recall that $\chi=\theta_{\text{meta}}-\alpha\nabla L_a(\theta_{\text{meta}})$ and $\chi^{*}=\arg\min L_{\tau}(\theta)=\theta_a^{*}-\alpha\nabla L_{\tau}(\theta_a^{*})$, where θ_{meta} can be considered as an estimation of θ_a^{*} and $L_a(\cdot)$ is the sample average approximation of $L_{\tau}(\cdot)$. In addition, denote $\tilde{\chi}=\theta_a^{*}-\alpha\nabla L_a(\theta_a^{*})$.

$$\begin{aligned}\|\chi-\chi^*\| &= \|\chi-\tilde{\chi}+\tilde{\chi}-\chi^*\| \\ &\le \|\chi-\tilde{\chi}\|+\|\tilde{\chi}-\chi^*\|\end{aligned} \tag{D1}$$

where $\|\chi-\tilde{\chi}\|$ denote the gap between the output model of RL θ_{meta} and the objective optimal model θ_a^*, and $\|\tilde{\chi}-\chi^*\|$ represents the error introduced by the sample average approximation of the loss function.

For $\|\chi-\tilde{\chi}\|$ in (D1), we can express the term as

$$\begin{aligned}\|\chi-\tilde{\chi}\| &= \left\|\theta_{\text{meta}}-\theta_a^*-\alpha\left[\nabla L_a\left(\theta_{\text{meta}}\right)-\nabla L_a\left(\theta_a^*\right)\right]\right\| \\ &\le \left\|\theta_{\text{meta}}-\theta_a^*\right\|+\alpha\left\|\nabla L_a\left(\theta_{\text{meta}}\right)-\nabla L_a\left(\theta_a^*\right)\right\| \\ &\le (1+\alpha H)\left\|\theta_{\text{meta}}-\theta_a^*\right\| \\ &= (1+\alpha H)\left\|\theta_{\text{meta}}-\theta^*+\theta^*-\theta_a^*\right\| \\ &\le (1+\alpha H)\left[\|\theta_{\text{meta}}-\theta^*\|+\left\|\theta^*-\theta_a^*\right\|\right] \\ &\le (1+\alpha H)\left(\zeta+\left\|\theta^*-\theta_a^*\right\|\right).\end{aligned} \tag{D2}$$

Next, we evaluate the term $\|\tilde{\chi}-\chi^*\|$ in (D1). Note that

$$\|\tilde{\chi}-\chi^*\| = \alpha\left\|\nabla L_\tau\left(\theta_a^*\right)-\nabla L_a\left(\theta_a^*\right)\right\| \tag{D3}$$

where $\nabla L_a(\cdot) = \frac{1}{k}\sum_{\mathcal{T}_x\in\mathcal{T}}\nabla l(\cdot,\mathcal{T}_x)$ and $\nabla L_\tau(\cdot) = \mathbb{E}_{\tau\sim p}\nabla l(\cdot,\theta)$.

Define $y \triangleq \nabla l(\cdot)$, then we have

$$Y_a\left(\theta_a^*\right) \triangleq \frac{1}{k}\sum\nolimits_{\mathcal{T}_x\in\mathcal{T}} y\left(\theta_a^*,\mathcal{T}_x\right) = \nabla L_a\left(\theta_a^*\right) \tag{D4}$$

and

$$Y_\tau\left(\theta_a^*\right) \triangleq \mathbb{E}_{\tau\sim p} y\left(\theta_a^*,\mathcal{T}_x\right) = \nabla L_\tau\left(\theta_a^*\right). \tag{D5}$$

Similarly, $Y_a(\cdot)$ is the sample average approximation of $Y_\tau(\cdot)$ and $l(\theta)$ is H-smooth. According to the uniform law of large number, there exist constants C_a and $n = n(\zeta)$ such that

$$\Pr\{\sup\|Y_a(\theta)-Y_\tau(\theta)\| \ge \zeta\} \le C_a \mathrm{e}^{-kn} \tag{D6}$$

where $\forall\zeta > 0$.

According to (D3), (D4), and (D5), we have

$$\Pr\{\|\tilde{\chi}-\chi^*\| \le \alpha\zeta\} \ge 1-C_a\mathrm{e}^{-kn}. \tag{D7}$$

Combining (D2) and (D7), we can derived the following expression

$$\begin{aligned}\|L_\tau(\chi)-L_\tau(\chi^*)\| &\le \alpha H\zeta + H(1+\alpha H)\zeta \\ &\quad + H(1+\alpha H)\|\theta^*-{\theta_a}^*\|\end{aligned} \tag{D8}$$

which illustrates the effect of task variability on the adaptation performance.