

Multi-User Collaborative Jamming Deception for UAV Communications: A Multi-Agent Reinforcement Learning Based Method

Guoyang Zhang, *Graduate Student Member, IEEE*, Zipeng Yang, *Graduate Student Member, IEEE*, Zhenyu Xiao, *Senior Member, IEEE*, Zhu Han, *Fellow, IEEE*, and Xiang-Gen Xia, *Fellow, IEEE*

Abstract—Due to the flexible mobility and high-quality of line-of-sight (LoS) channels, unmanned aerial vehicle (UAV) has begun to play an important role in wireless communications. However, the broadcasting nature of wireless communications and the limited payload of UAVs render the spectrum vulnerable to malicious jamming attacks. To guarantee the performance of UAV communications, this paper focuses on reactive jamming, and sets up an actively exposed deception band to attract partial power of jamming. Specifically, we first model the anti-jamming process with a Stackelberg game model, under the assumption that the rational behavior of jamming is known. Then, we analyze the theoretical optimal strategies of the jamming as well as the users in UAV communication to reach the equilibrium of the above game model. Finally, we design a collaborative multi-agent jamming deception method to achieve anti-jamming in the absence of environmental and jamming information. This method is based on the centralized evaluation network at the UAV and decentralized policy network at each user. Simulation results show that the anti-jamming performance of the proposed method can approach the theoretical upper bound and significantly outperform other benchmark methods.

Index Terms—UAV communications, jamming deception, actor critic, multi-agent reinforcement learning, multi-user collaborative.

I. INTRODUCTION

WITH the advantages of low cost, flexible deployment, and high robustness of line-of-sight (LoS) channels, unmanned aerial vehicles (UAVs) have been considered as promising devices to assist wireless communications [1]–[3]. However, due to the broadcasting nature of wireless communications, the signal transmission can be easily monitored by malicious users (MUs). As a result, the spectrum used by legitimate users (LUs) can be potentially attacked by MUs [4]–[6]. Furthermore, as communication devices with limited payload capacity, UAVs are more severely affected by spectrum jamming compared to terrestrial facilities [7]–[10], which necessitates the research on anti-jamming methods to guarantee the performance of UAV-assisted communications.

Guoyang Zhang, Zipeng Yang, and Zhenyu Xiao are with the School of Electronic and Information Engineering, Beihang University, Beijing 100191, China. (e-mail: {zhangguoyang, yangzipeng, xiaozy}@buaa.edu.cn). The corresponding author is Dr. Zhenyu Xiao with Email xiaozy@buaa.edu.cn.

Zhu Han is with the Department of Electrical and Computer Engineering at the University of Houston, Houston, TX 77004 USA, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul, South Korea, 446-701. (e-mail: hanzhu22@gmail.com).

Xiang-Gen Xia is with the Department of Electrical and Computer Engineering, University of Delaware, Newark, DE 19716, USA (e-mail: xxia@ee.udel.edu).

The traditional spread spectrum techniques, such as direct sequence [11]–[14] and frequency hopping [15]–[17], consume extra spectrum resources. Other anti-jamming technologies, such as time hopping [18] and multiple antennas [19], require pre-designed passive anti-jamming strategies, thus resulting in low efficiency in the utilization of available resources and poor adaptability. Thus, these anti-jamming technologies sacrifice communication performance and increase hardware costs, whose effectiveness is limited in payload-constrained UAV communications.

With the development of spectrum sensing and software radio technologies, perception-guided intelligent jamming techniques increasingly pose a severe threat to UAV communications [10], [20]. As a result, the advancement of intelligent anti-jamming technologies have been promoted. To effectively characterize the anti-jamming process and formulate anti-jamming decisions, many researchers have modeled the interaction between communication systems and jamming. Typical modeling methods include the Markov decision process (MDP) and game theory. MDP, in particular, is a method for modeling the dynamic interaction process between agents and the environment [20]–[22]. It consists of states, actions, transition probabilities, and rewards. Strategies are updated through rewards obtained from interactions with the environment to maximize the expected cumulative reward. Common solutions include reinforcement learning [23] and swarm intelligence algorithms [24]. Due to the independence from historical information and the ability to make optimal decisions based on the current state, MDP has received considerable attention from researchers. For example, the authors in [21] modeled the interaction between UAV wireless communications and intelligent jamming with an MDP and designed a solution based on deep reinforcement learning. Additionally, the authors in [25] modeled the counteractive process of reactive jamming as an MDP and employed reinforcement learning to attract single-band reactive jamming to a specific frequency band.

Game theory models the performance of UAV communication systems, jamming, and user actions as a zero-sum game to study the optimal decisions between different parties [26]–[30]. When model parameters are fully known, researchers attempt to reach the game equilibrium. When parameters are unknown, learning algorithms are commonly employed. Common anti-jamming games include communication anti-jamming games based on frequency hopping and power allocation anti-jamming games. For instance, the authors in [31]

used the Stackelberg game model to model the behaviors of LUs and MUs, where LUs design strategies to coordinate transmission power and anti-jamming performance. The Stackelberg game provides a natural framework for sequential decision-making in security and anti-jamming contexts, with seminal applications in communication networks established in early works such as [27]. The authors in [32] employed a bimatrix game model to model the anti-jamming process. The authors derived the Nash equilibrium of the game with known global information and developed a method to approach Nash equilibrium when information is unknown. As the number of users increases and the network scale expands, the action set of users in the game model will face the problem of rapidly increasing dimension. As a result, the anti-jamming decisions become difficult to converge to a feasible solution [33]–[35]. Furthermore, information sharing between LUs is difficult to be guaranteed due to the existence of reactive jamming. Therefore, it is not realistic for each user to make decisions based on global information [36]. To this end, the authors in [37] proposed an anti-jamming method based on multi-agent reinforcement learning (MARL), which employs policy network on each LU, and makes decisions based on the local observations of each LU. With the help of MARL, the complexity of the solution to anti-jamming could be significantly reduced. Also, each LU can overcome the dependence on global information and make anti-jamming decisions with its own policy network. However, considering only local observations of each LU may lead to spectrum competition conflicts and loss of global optimal performance.

In summary, the existing solutions primarily counteract jamming through high-dimensional global decisions, which are challenging to be applied in payload-limited UAV communications. While conventional optimization-based methods rely on perfect yet often unavailable global channel information, incurring prohibitive communication and computation overhead, and rule-based strategies lack the adaptability to counter intelligent reactive jammers, most existing learning-based approaches still face the curse of dimensionality in decision-making.

Our work bridges two distinct research strands to address these limitations: 1) game-theoretic modeling for anti-jamming strategy analysis, and 2) scalable multi-agent learning for decentralized execution. On the theoretical modeling front, Stackelberg games have been established as a principled framework for analyzing hierarchical interactions between defenders and attackers in communication security, capturing the sequential nature of reactive jamming [27]–[29]. However, such game-theoretic solutions typically assume complete information and centralized computation, limiting their applicability in practical UAV scenarios with partial observability and computational constraints. On the algorithmic implementation front, the Centralized Training with Decentralized Execution (CTDE) paradigm has emerged as a powerful MARL framework for enabling scalable coordination in complex wireless networks [38]–[40]. Yet, existing CTDE applications in wireless security often lack strong theoretical foundations regarding optimal strategy design.

To bridge this gap, we propose a collaborative multi-

agent jamming deception (CMAJD) method that integrates Stackelberg game-theoretic modeling with a CTDE-based MARL framework specifically tailored for UAV anti-jamming. Given the above considerations, this paper focuses on the challenge of computationally load-constrained UAV that is unable to handle high-dimensional anti-jamming problems in wireless communications. Our CMAJD method first employs a Stackelberg game to derive theoretical optimal strategies under complete information, establishing a performance upper bound. Then, we design a CTDE-based MARL algorithm that enables decentralized agents to learn near-optimal strategies in unknown and dynamic environments. With the help of decentralized policy networks, the high-dimensional global decision space is decomposed into low-dimensional local decisions per user, achieving a fundamental reduction in computational complexity (a quantitative analysis is provided in Section III-D). This reduction enables real-time anti-jamming decisions on payload-constrained UAVs. Moreover, the global performance of these local decisions is guided and ensured by the centralized value network at the UAV. To clearly position our contributions relative to existing approaches, we provide a comparative summary in Table I.

The main contributions of this paper are summarized as follows:

- 1) We investigate a multi-band reactive jamming in UAV assisted wireless communications, modeling the interaction between LUs and an MU as a multi-leader single-follower Stackelberg game model, where LUs are leaders and MU is the follower. This model captures the sequential decision-making inherent to reactive jamming.
- 2) We derive the band selection and power allocation to achieve the game equilibrium between LUs and MU when all environmental information is known, establishing a theoretical performance upper bound. Then, we design an MDP to depict the anti-jamming process. A CMAJD method is designed based on the CTDE framework, considering the difficulty of obtaining complete environmental information and global states in UAV communications.
- 3) The proposed method sets a deception band that could be deliberately exposed to attract the power of MU, acting as a sacrificial lamb to protect genuine communications. Among all LUs, we deploy policy networks in a distributed manner. Each LU can make decision according to the policy network based on its own local observation (e.g., sensed jamming power). Thus, the dimension of global anti-jamming decision can be reduced, which incurs lower computational overhead for real-time operation. At the UAV end, a centralized value network is deployed to assess the impact of distributed decisions on global performance, guiding the update of policy networks, maintaining the anti-jamming performance of low-dimensional collaborative decision.
- 4) Numerical results demonstrate that the proposed CMAJD method has good convergence and can effectively allocate communication bands and deception

TABLE I
COMPARISON WITH RELATED WORKS

Work	System Model	Learning Framework	Action Space	Deception Strategy
[21]	UAV cellular	DRL (single-agent)	Discrete	None
[25]	Single-user, single-band	DDPG	Continuous	Fixed band
[30]	Multi-user	Game theory	Continuous	None
[37]	Multi-user, multi-band	Independent MARL	Discrete	None
[41]	Multi-user, multi-tone	Centralized RL	Discrete	Multi-band
Ours	Multi-user, multi-band	CTDE-MARL	Hybrid	Collaborative

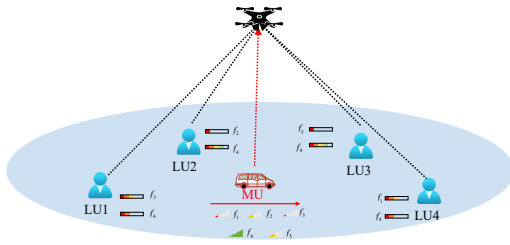


Fig. 1. System model of UAV-assisted wireless communication with a dynamic reactive jamming.

bands to avoid interference among LUs. It also shows strong adaptability to dynamic jamming strategies. The anti-jamming performance of CMAJD can approach the theoretical upper bound, and is superior to benchmarks in terms of system throughput.

The rest of this paper is organized as follows. In Section II, we introduce the system model and formulate the game model for anti-jamming. In Section III, we detail our proposed method. In Section IV, the simulation results are provided to show the superiority of the proposed method. Finally, in Section V we conclude this paper.

II. SYSTEM MODEL AND PROBLEM FORMULATION

As illustrated in Fig. 1, this paper considers a UAV-assisted uplink wireless communication system. The UAV is positioned at $\omega^u = [x^u, y^u, h]$, offering L available frequency bands for uplink transmission, denoted by $\mathcal{L} = \{1, 2, \dots, L\}$. There are N LUs on the ground, denoted by $\mathcal{N} = \{1, 2, \dots, N\}$, with the position of the n -th LU given by $\omega_n^{lu} = [x_n^{lu}, y_n^{lu}]$. It is assumed that each LU constantly has data packets to transmit, with a maximum transmitting power of P_{\max} . To guarantee signal transmission, the total spectrum of UAV is sufficiently divided for LU transmission, i.e., $L > N$. However, a dynamic MU performing reactive jamming also exists on the ground, with its position in each time slot given by $\omega_t^{mu} = [x_t^{mu}, y_t^{mu}]$, where $t = 1, 2, \dots, T$. The color of the signal strength bars in Fig. 1 represents the magnitude of the power. In each time slot, the MU senses the transmission power of LUs in each frequency band and allocates jamming power based on the perceived power, thereby disrupting LU communication. To ensure normal communication while combating jamming, in each time slot, each LU divides its transmission power into

TABLE II
MAIN KEY MATH SYMBOL DEFINITIONS

Symbol	Physical Meaning
\bar{g}^{g2a}	Channel gain between ground user and the UAV
θ	Elevation angle between one user and the UAV
$P_r^{LoS}(\theta)$	Probability of LoS link
$P_r^{NLoS}(\theta)$	Probability of non-line of sight (NLoS) link
$\bar{g}^{LoS}, \bar{g}^{NLoS}$	Channel gains for LoS and NLoS links between the user and the UAV
d_{g2a}	Euclidean distance for communication between the user and the UAV
\bar{g}^{g2g}	Channel gain between LU and MU
d_{g2g}	Euclidean distance between LU and MU
ς	Channel fading factor
P_{L2M}, P'_{L2M}	Power perceived by the MU in the communication and deception band
$x_n(t), x'_n(t)$	Communication and deception power of the n -th LU at timeslot t
$y_n(t), y'_n(t)$	Jamming power allocated to the communication and deception bands of the n -th LU by the MU
$l_n^C(t), l_n^D(t)$	communication and deception bands selected by the n -th LU at timeslot t
$y_{l_n^C}(t)$	Jamming power allocated by MU to frequency band l_n^C at timeslot t
$y_D(t)$	Total jamming power allocated by the MU to the deception band

two frequency bands among the L bands. One band is used for normal communication and the other one is employed to deceive the MU to waste its partial power to the deception band. Therefore, the jamming power in the communication band can be reduced. Moreover, LUs need to select different communication bands to avoid intra-user interference, while opting for the same deception band to conserve spectrum resources. The definitions of other key symbols in Section II are shown in Table II.

A. Channel Model

First, the channel gain between a ground user (either LU or MU) and the UAV is given by [42]

$$\bar{g}^{g2a} = \text{Pr}^{LoS}(\theta)\bar{g}^{LoS} + \text{Pr}^{NLoS}(\theta)\bar{g}^{NLoS}, \quad (1)$$

where θ denotes the elevation angle between one user and the UAV, determining the probability of an LoS link occurring between the user and the UAV [43], i.e.,

$$\Pr^{LoS}(\theta) = \frac{1}{1 + a \exp(-b(\theta - a))}, \quad (2)$$

where a and b represent the modeling parameters related to the environment. Then, the probability of a NLoS link occurring can be given by

$$\Pr^{NLoS}(\theta) = 1 - \Pr^{LoS}(\theta). \quad (3)$$

Furthermore, \bar{g}^{LoS} and \bar{g}^{NLoS} represent the channel gains for LoS and NLoS links between the user and the UAV, respectively, as given by

$$\bar{g}^{LoS} = \frac{c}{(4\pi f_l d_{g2a})^{\alpha_{g2a}}} \eta^{LoS}, \quad (4)$$

$$\bar{g}^{NLoS} = \frac{c}{(4\pi f_l d_{g2a})^{\alpha_{g2a}}} \eta^{NLoS}, \quad (5)$$

where $d_{g2a} = \sqrt{(x^u - x)^2 + (y^u - y)^2 + h^2}$ is the Euclidean distance for communication between the user and the UAV, h is the height of the UAV, c is the speed of electromagnetic wave propagation in free space, f_l is the center frequency of the l -th frequency band, α_{g2a} is the path loss parameter for the ground-to-air link, and η^{LoS} and η^{NLoS} are the large-scale fading for LoS and NLoS links, respectively.

Next, the channel gain between LU and MU is given by

$$\bar{g}^{g2g} = \frac{c}{(4\pi f_l d_{g2g})^{\alpha_{g2g}}} \varsigma, \quad (6)$$

where d_{g2g} represents the Euclidean distance between LU and MU, α_{g2g} is the path loss factor, and ς is the channel fading factor following an exponential distribution with unit mean.

B. Jamming Model

During the transmission process, the MU monitors the transmission power of each LU within a selected frequency band. At timeslot t , the power perceived by the MU in the communication band is given by

$$P_{L2M} = \sum_{n=1}^N x_n(t) \bar{g}_n^{g2g}(t), \quad (7)$$

where $x_n(t)$ is the communication power of the n -th LU at timeslot t , and $\bar{g}_n^{g2g}(t)$ is the channel gain of the communication band selected by the n -th LU at timeslot t . Similarly, the power perceived by the MU in the deception band is

$$P'_{L2M} = \sum_{n=1}^N x'_n(t) \bar{g}'_n^{g2g}(t), \quad (8)$$

where $x'_n(t)$ is the deception power of user n at timeslot t , and $\bar{g}'_n^{g2g}(t)$ is the channel gain of the deception band selected by the n -th LU at timeslot t . It is important to note that the MU can only perform power sensing. In other words, it cannot demodulate the signals to distinguish between the jamming and deception power. It can only allocate higher interference

power to the frequency band where higher power is detected. Therefore, the utility function of the MU can be defined as [41].

$$U_{MU} = \min_{\{y_n(t)\}, \{y'_n(t)\}} E \left\{ \sum_{n=1}^N \left(\frac{P_{L2M}}{y_n(t)} + \frac{P'_{L2M}}{y'_n(t)} \right) \right\}, \quad (9)$$

$$\text{s.t. } y_n(t), y'_n(t) \geq 0, \forall n \in N, \quad (9a)$$

$$\sum_{n=1}^N (y_n(t) + y'_n(t)) = P_{\max}^J, \quad (9b)$$

where $y_n(t)$ and $y'_n(t)$ represent the jamming power allocated to the communication and deception bands of the n -th LU by the MU, and P_{\max}^J is the maximum power of MU, respectively. The expectation in (9) captures the long-term average utility of the MU over the sequential game, accounting for the uncertainty in future states and actions.

C. Problem Formulation

As for the UAV-assisted wireless communication, the objective should be maximizing the signal to interference plus noise ratio (SINR) at the UAV end. Therefore, LUs need to select different communication frequency bands to avoid mutual interference. This paper aims to optimize the frequency band selection and power allocation for the LUs to maximize the expected SINR. Thus, the optimization problem is formulated as

$$U_{UAV} = \max_{\{x_n(t)\}, \{x'_n(t)\}, \{l_n^C(t)\}, \{l_n^D(t)\}} E \left\{ \sum_{n=1}^N \frac{x_n(t) \bar{g}_n^{g2a}(t)}{y_{l_n^C}(t) \bar{g}_J^{g2a}(t) + \sigma^2} \right\}, \quad (10)$$

$$\text{s.t. } x_n(t), x'_n(t) \geq 0, \forall n \in \mathcal{N}, \quad (10a)$$

$$x_n(t) + x'_n(t) = P_{\max}, \forall n \in \mathcal{N}, \quad (10b)$$

$$l_n^C(t) \neq l_m^D(t), \forall n \neq m, n, m \in \mathcal{N}, \quad (10c)$$

$$l_n^C(t) \neq l_m^C(t), \forall n \neq m, n, m \in \mathcal{N}, \quad (10d)$$

$$l_n^D(t) = l_m^D(t), \forall n \neq m, n, m \in \mathcal{N}, \quad (10e)$$

where σ^2 is the power of the additive white Gaussian noise. $l_n^C(t)$ and $l_n^D(t)$ represent the communication and deception bands selected by the n -th LU at timeslot t , respectively. P_{\max} is the maximum power of LUs, $y_{l_n^C}(t)$ denotes the total jamming power on the communication band $l_n^C(t)$ selected by LU n . This term captures the band-dependent nature of jamming interference. The value of $y_{l_n^C}(t)$ is determined by the jamming power allocations of all LUs that use the same band, either for communication or deception:

$$y_{l_n^C}(t) = \sum_{m=1}^N [\delta(l_n^C(t) - l_m^C(t)) y_m(t) + \delta(l_n^C(t) - l_m^D(t)) y'_m(t)], \quad (11)$$

where $\delta(\cdot)$ is the Kronecker delta function ($\delta(0) = 1$, $\delta(x) = 0$ for $x \neq 0$). In the context of optimization problem (10), $y_{l_n^C}(t)$ is a determined quantity that depends on the band selections $l_n^C(t)$ and $l_n^D(t)$ from the previous timeslot (or the current timeslot if considering a sequential decision process),

rather than being a function of the optimization variables in (10) itself. Constraints (10a) and (10b) confine the feasible region of power used for communication and deception. Constraint (10c) ensures that the signal transmission of each LU is not influenced by the deception activities of other LUs. Constraint (10d) means that different LUs should choose distinct communication bands to avoid interference during signal transmission. Lastly, constraint (10e) implies that LUs should select the same band for deception to prevent spectrum wastage.

D. Modeling the Stackelberg Game Model

In each timeslot, all LUs first choose appropriate communication and deception bands and allocate power accordingly. As a reactive jamming with intelligent and rational characteristics [44], MU will allocate its own power based on the sensing results to achieve optimal jamming performance. Therefore, the MU monitors the power in each band and injects jamming power accordingly, aiming to disrupt the signal transmission of the LUs. Finally, the UAV evaluates the SINR from each LU to assess the communication quality of the entire system in each time slot. In this process, the interaction between the LUs and the MU can be modeled as a Stackelberg game model, which is a non-cooperative game model consisting of leaders and followers. In the scenario considered in this paper, the LUs are considered as multiple leaders, and the MU is the follower. The specific game model is defined as:

$$\mathcal{G}(t) = \{\{\text{LUs, MU}\}, \{\mathcal{A}_{LU}(t), \mathcal{A}_{MU}(t)\}, \{U_{UAV}, U_{MU}\}\}, \quad (12)$$

where $\mathcal{A}_{LU}(t) = \{x_n(t), x'_n(t), l_n^C(t), l_n^D(t)\}$ represents the action set of the LUs, indicating the communication power, the deception power, the communication bands and the deception bands for each LU. $\mathcal{A}_{MU}(t) = \{y_n(t), y'_n(t)\}$ is the action set of the MU, representing the jamming power allocated by the MU on the communication and deception bands of the LUs. During the game, the LUs and the MU continuously adjust and optimize $\mathcal{A}_{LU}(t)$ and $\mathcal{A}_{MU}(t)$ to maximize the utility functions of (10) and (9), respectively.

III. PROPOSED METHOD

In this section, the method for maximizing the utility function in (10) will be introduced. Initially, assuming that all relevant information is known, an analysis of (12) is conducted to derive the equilibrium solution to all the LUs and the MU. Consequently, the optimal solution to (10) can be obtained by giving the known channel information and jamming power. However, it is challenging for the LUs to ascertain the behavior of the MU, and the channel information becomes uncertain due to the dynamic nature of the MU. To address this challenge, a method based on multi-agent reinforcement learning is designed to deceive the MU when environmental information is unknown.

A. Analysis for Optimal Solution

In any given timeslot, assuming $\mathcal{A}_{LU}(t)$ is known to the MU, the function depicted in (9) can be characterized as

concave, which implies the certainty of a unique optimal solution. Analogously, with $\mathcal{A}_{MU}(t)$ known to the LU, the linearity of the function in (10) similarly guarantees the identification of a unique optimal solution. Therefore, when the specific bands are selected, both LUs and MU have optimal actions, and would not change their strategy voluntarily, which leads to a game equilibrium in (12). In this subsection, we will first analyze the equilibrium point of the game.

According to (10e), all the LUs will choose the same band for deception. To analyze the equilibrium at a single time slot, we consider the instantaneous optimization problem with known channel states. Thus, the expectation is removed. Additionally, since all LUs share the same deception band (constraint (10e)), the jamming power on that band can be aggregated into $y_D(t)$. The problem then becomes:

$$U_{MU} = \min_{\substack{\{y_n(t)\}, \\ \{y_D(t)\}}} \left(\sum_{n=1}^N \frac{x_n(t) \bar{g}_n^{g2g}(t)}{y_n(t)} + \frac{\sum_{n=1}^N x'_n(t) \bar{g}'_n{}^{g2g}(t)}{y_D(t)} \right), \quad (13)$$

$$\text{s.t. } y_n(t), y_D(t) \geq 0, \forall n \in N, \quad (13a)$$

$$y_D(t) + \sum_{n=1}^N y_n(t) = P_{\max}^J. \quad (13b)$$

Construct the Lagrangian function for (13) to determine the optimal solution:

$$\begin{aligned} L(y_n(t), y_D(t), \lambda^J, \mu^J) &= \sum_{n=1}^N \frac{x_n(t) \bar{g}_n^{g2g}(t)}{y_n(t)} \\ &+ \frac{\sum_{n=1}^N x'_n(t) \bar{g}'_n{}^{g2g}(t)}{y_D(t)} \\ &- \sum_{n=1}^N \lambda_n^J y_n(t) - \lambda_D^J y_D(t) \\ &+ \mu^J \left(\sum_{n=1}^N y_n(t) + y_D(t) - P_{\max}^J \right), \end{aligned} \quad (14)$$

where λ_n^J , λ_D^J , and μ^J are Lagrange multipliers. By applying the Karush-Kuhn-Tucker (KKT) conditions to the above equation, we can derive:

$$\frac{\partial L}{\partial y_n(t)} = -\frac{x_n(t) \bar{g}_n^{g2g}(t)}{(y_n(t))^2} - \lambda_n^J + \mu^J = 0, \quad (15a)$$

$$\frac{\partial L}{\partial y_D(t)} = -\frac{\sum_{n=1}^N x'_n(t) \bar{g}'_n{}^{g2g}(t)}{(y_D(t))^2} - \lambda_D^J + \mu^J = 0, \quad (15b)$$

$$\lambda^J y_n(t) = 0, \quad (15c)$$

$$\frac{\partial L}{\partial \mu^J} = \sum_{n=1}^N y_n(t) + y_D(t) - P_{\max}^J = 0. \quad (15d)$$

In the KKT conditions for the MU's optimization problem (13), the variables $y_n(t)$ and $y_D(t)$ appear in the denominator of terms in equations (15a) and (15b). Therefore, for the

KKT conditions to be well-defined, we must have $y_n(t) > 0$ and $y_D(t) > 0$. Consequently, the non-negativity constraints $y_n(t) \geq 0$ and $y_D(t) \geq 0$ are not binding at the optimum, and from the complementary slackness condition (15c), we obtain $\lambda^J = 0$. From (15a), $y_n(t) = \sqrt{\frac{x_n(t)\bar{g}_n^{g2g}(t)}{\mu^J}}$ can be derived. According to (15b), we can find that $y_D(t) = \sqrt{\frac{\sum_{n=1}^N x'_n(t)\bar{g}'_n{}^{g2g}(t)}{\mu^J}}$. Substituting the aforementioned results into (15d), we can deduce:

$$\sqrt{\mu^J} = \frac{\sum_{n=1}^N \sqrt{x_n(t)\bar{g}_n^{g2g}(t)} + \sqrt{\sum_{n=1}^N x'_n(t)\bar{g}'_n{}^{g2g}(t)}}{P_{\max}^J}, \quad (16)$$

Therefore, the optimal solution of the MU in each timeslot can be obtained as

$$y_n(t) = \frac{P_{\max}^J \sqrt{x_n(t)\bar{g}_n^{g2g}(t)}}{\sum_{n=1}^N \sqrt{x_n(t)\bar{g}_n^{g2g}(t)} + \sqrt{\sum_{n=1}^N x'_n(t)\bar{g}'_n{}^{g2g}(t)}}, \quad (17)$$

$$y_D(t) = \frac{P_{\max}^J \sqrt{\sum_{n=1}^N x'_n(t)\bar{g}'_n{}^{g2g}(t)}}{\sum_{n=1}^N \sqrt{x_n(t)\bar{g}_n^{g2g}(t)} + \sqrt{\sum_{n=1}^N x'_n(t)\bar{g}'_n{}^{g2g}(t)}}. \quad (18)$$

Under the Stackelberg game equilibrium, all LUs select distinct communication bands and share a common deception band, i.e., constraints (10d) and (10e) are satisfied. Consequently, for each LU n , the band-specific jamming power $y_{jC}_n(t)$ equals the jamming power $y_n(t)$ allocated to its communication band, as derived in Section III-A. Therefore, we can substitute $y_{jC}_n(t)$ with $y_n(t)$ in (10). However, even with this substitution, (10) will be extremely difficult to be solved with the results of (17). In addition, the noise power is small and negligible relative to the jamming power. Therefore, we follow the descriptions in [25] and [41] and ignore the noise term in (10), taking the maximization of the signal-to-interference ratio (SIR) as the objective. This is justified in high-jamming scenarios where interference power dominates over noise. Substituting the optimal jamming powers $y_n(t)$ from (17) into (10) yields,

$$U_{UAV} = \max_{x_n(t), x'_n(t)} \sum_{n=1}^N \phi_n \sqrt{x_n(t)} \times \left(\sum_{n=1}^N \sqrt{x_n(t)\bar{g}_n^{g2g}(t)} + \sqrt{\sum_{n=1}^N x'_n(t)\bar{g}'_n{}^{g2g}(t)} \right), \quad (19)$$

$$\text{s.t. } x_n(t), x'_n(t) \geq 0, \forall n \in N, \quad (19a)$$

$$x_n(t) + x'_n(t) \leq P_{\max}, \quad (19b)$$

where $\phi_n = \frac{\bar{g}_n^{g2g}}{P_{\max}^J \bar{g}_n^{g2g}(t) \sqrt{\bar{g}_n^{g2g}(t)}}$. By computing the Hessian

matrix of (19), it can be demonstrated that the function is concave [45]. Similarly, when the behavior of the MU is given, the unique solution to (19) can be determined through the application of the Lagrange multiplier method. The Lagrangian function can be constructed as:

$$L(x_n(t), x'_n(t), \lambda, \mu) = - \sum_{n=1}^N \phi_n \sqrt{x_n(t)} \times \left(\sum_{n=1}^N \sqrt{x_n(t)\bar{g}_n^{g2g}(t)} + \sqrt{\sum_{n=1}^N x'_n(t)\bar{g}'_n{}^{g2g}(t)} \right) - \lambda_n x_n(t) - \lambda_{N+n} x'_n(t) + \mu_n (x_n(t) + x'_n(t) - P_{\max}), \quad (20)$$

where λ_n , μ_n , and λ_{N+n} are the Lagrange multipliers, applying the KKT conditions to (20), we can obtain:

$$\frac{\partial L}{\partial x_n(t)} = - \frac{\phi_n}{2\sqrt{x_n(t)}} \left(\sum_{n=1}^N \sqrt{x_n(t)\bar{g}_n^{g2g}(t)} + \sqrt{\sum_{n=1}^N x'_n(t)\bar{g}'_n{}^{g2g}(t)} \right) - \frac{N}{2} \phi_n \sqrt{\bar{g}_n^{g2g}(t)} - \lambda_n + \mu_n = 0, \quad (21a)$$

$$\frac{\partial L}{\partial x'_n(t)} = - \sum_{n=1}^N \frac{\phi_n \bar{g}'_n{}^{g2g}(t) \sqrt{x_n}}{2\sqrt{\sum_{n=1}^N x'_n \bar{g}'_n{}^{g2g}(t)}} - \lambda_{N+n} + \mu_n = 0, \quad (21b)$$

$$\lambda_n x_n(t) = 0, \quad (21c)$$

$$\lambda_{N+n} x'_n(t) = 0, \quad (21d)$$

$$\mu_n (x_n(t) + x'_n(t) - P_{\max}) = 0. \quad (21e)$$

For the KKT conditions (21a)-(21b) to be well-defined, the terms $1/\sqrt{x_n(t)}$ and $1/\sqrt{\sum_{n=1}^N x'_n(t)\bar{g}'_n{}^{g2g}(t)}$ must be finite. This necessitates $x_n(t) > 0$ and $\sum_{n=1}^N x'_n(t)\bar{g}'_n{}^{g2g}(t) > 0$. In the cooperative equilibrium, we assume $x'_n(t) > 0$ for all n . Therefore, the constraints $x_n(t) \geq 0$ and $x'_n(t) \geq 0$ are not binding, yielding $\lambda_n = 0$, $\lambda_{N+n} = 0$ from (21c)-(21d). By substituting these results and $x_n(t) = P_{\max} - x'_n(t)$ into (21a) and (21b), we can obtain:

$$\frac{\phi_n}{\sqrt{x_n}} \left(\sum_{n=1}^N \sqrt{x_n \bar{g}_n^{g2g}(t)} + \sqrt{\sum_{n=1}^N (P_{\max} - x_n) \bar{g}'_n{}^{g2g}(t)} \right) + N \phi_n \sqrt{\bar{g}_n^{g2g}(t)} - \sum_{n=1}^N \frac{\phi_n \bar{g}'_n{}^{g2g}(t) \sqrt{x_n}}{\sqrt{\sum_{n=1}^N (P_{\max} - x_n) \bar{g}'_n{}^{g2g}(t)}} = 0. \quad (22)$$

Note that directly solving (22) is challenging, but the numerical solutions can be obtained through simulation software to determine the optimal response of the UAV-assisted communication system to the MU.

Combining (22), (17), and (18), the optimal strategies for the LUs can be derived with fully known channel information and behavior of the MU. However, we have to enumerate and sort through all possible selections of frequency bands to obtain

$$\text{SINR}_n(t) = \frac{x_n(t)\bar{g}_n^{g^{2a}}(t)}{y_{l_n^C}(t)\bar{g}_J^{g^{2a}}(t) + \sum_{m=1, m \neq n}^N \left[\delta(l_n^C(t) - l_m^C(t))x_m(t)\bar{g}_m^{g^{2a}}(t) + \delta(l_n^C(t) - l_m^D(t))x'_m(t)\bar{g}_m^{g^{2a}}(t) \right] + \sigma^2}, \quad (23)$$

the optimal strategy across all bands. For instance, assuming there are $L = 10$ frequency bands and $N = 8$ users, the total number of combinations to be enumerated is $10!$. As the number of frequency bands and users increases, the number of combinations to enumerate rises super exponentially. Thus, even with complete environment information available, finding the optimal solution with the above method still faces the challenge of excessive complexity. Moreover, in the presence of jamming, each LU cannot obtain the environment information of other LUs and must make decision based solely on its own observations, making it impossible to achieve a globally optimal strategy for jamming deception.

B. Set Up MDP

The previously described method poses high demands on the computational capabilities of UAV-assisted wireless communications. To this end, in this subsection, a jamming deception method based on MARL is designed, i.e., CMAJD. Notably, in this paper, the location of the MU dynamically changes. In this regard, the channel gains between the MU and LUs, as well as between the MU and UAV, vary over time. Moreover, the current position of the MU influences the channel states in the subsequent moment. Therefore, the dynamic interaction between the UAV communication system and the MU can be effectively modeled using an MDP.

The MDP is described by a tuple $\langle \mathcal{S}(t), \mathcal{O}(t), \mathcal{A}, \mathcal{R} \rangle$, where $\mathcal{S}(t)$ denotes the global state observed by the UAV. $\mathcal{O}(t) = \{o_1(t), o_2(t), \dots, o_N(t)\}$ represents the local observations of the LUs. \mathcal{A} and \mathcal{R} in the tuple represent actions of all LUs and the reward obtained from the environment after performing actions, respectively. In each timeslot, the action of each LU will be made based on its own observation. To preserve the optimality of decision-making, we design a hybrid action set that combines discrete and continuous actions, i.e., $a_n(t) = \{a_n^{dis}(t), a_n^{con}(t)\}$. Specifically, discrete action $a_n^{dis}(t)$ means that each LU needs to choose a deception band and a communication band in the available bands whose center frequency is discretized. Continuous action $a_n^{con}(t)$ means that each LU needs to perform a continuous allocation of deception power and communication power. Different from the setting in [41] that directly discretizes the power to $\chi + 1$ level, we perform continuous power allocation to avoid the loss of optimality caused by discretization. By decomposing the high-dimensional global action space into distributed low-dimensional decisions per user, CMAJD significantly reduces computational complexity. A detailed quantitative complexity analysis is provided in Section III-D. From the collective actions $\mathcal{A}(t) = \{a_1(t), a_2(t), \dots, a_N(t)\}$ of all LUs, UAV can determine the global state $\mathcal{S}(t)$ and current reward \mathcal{R} can be calculated based on $\mathcal{S}(t)$. The detailed definition of the MDP is as follows.

1) *Global State $\mathcal{S}(t)$* : The utility function of the UAV communication system defined in (10) aims to maximize the expected SINR while avoiding interference caused by the communication power and deception power between the LUs. Therefore, in the global state, the SINR of each LU needs to be calculated at first by (23) at the top of the next page. Note that equation (11) uses $y_{l_n^C}(t)$ instead of $y_n(t)$ because it describes the interference during the game process, where the system may not have reached the equilibrium point. In (10), the numerator represents the received power at the UAV from the n -th LU, while the denominator comprises two parts: the first part accounts for the jamming caused by MU, and the second part represents the co-channel interference among LUs. We note that in a practical UAV-assisted system, the UAV cannot directly observe the MU's location $\omega^{mu}(t)$ or instantaneous channel gains. However, the selected state variables $\text{SINR}(t)$, $k^C(t)$, and $k^D(t)$ are sufficient to form a Markov state for the following reasons: 1) $\text{SINR}(t)$ inherently captures the aggregated impact of MU's jamming strategy, channel conditions, and mobility on the system performance at time t ; 2) $k^C(t)$ and $k^D(t)$ summarize the current band allocation outcome, which is a result of all agents' past actions and environmental interactions. These variables collectively provide a compressed representation of the history that is relevant for future decision-making, satisfying the Markov property in the context of our decentralized control problem. Finally, the global state is defined as:

$$\mathcal{S}(t) = \{\text{SINR}(t), k^C(t), k^D(t)\}, \quad (24)$$

where $\text{SINR}(t) \triangleq \{\text{SINR}_n(t) | n = 1, 2, \dots, N\}$.

2) *Local Observations $\mathcal{O}(t)$* : In scenarios with reactive jamming, it becomes challenging for LUs to share information. Consequently, each LU can only observe its own state and make decisions based on this local insight. The local observation available to each LU is given by:

$$o_n(t) = \{x_n(t), x'_n(t), l_n^C(t), l_n^D(t)\}, \quad (25)$$

which includes the agent's own transmitted power and selected bands. This design reflects practical constraints in UAV-assisted systems: each ground user has limited sensing capabilities and cannot reliably obtain real-time measurements of environmental variables (e.g., instantaneous interference or channel gains) under jamming. Although $o_n(t)$ does not contain direct environmental feedback, it provides a *minimal yet sufficient* basis for decentralized decision-making within the CTDE framework. The centralized critic at the UAV, which has access to the global state $\mathcal{S}(t)$, evaluates the joint effect of all agents' actions and provides the necessary guidance

during training. This decoupling allows each LU to focus on its own controllable decisions while the critic ensures global coordination. The empirical results in Section IV demonstrate that this observation structure, coupled with the CTDE architecture, enables the agents to learn effective cooperative policies that achieve near-optimal performance.

Aggregating the local observations of all LUs, the global state $\mathcal{S}(t)$ can be obtained.

- 3) *Hybrid Action* $\mathcal{A}(t)$: In each time slot, each LU will choose bands for communication and deception, allocate communication power, and distribute the remaining power to the deception band. Therefore, the action of each LU can be defined as:

$$a_n(t) = \{l_n^C(t), l_n^D(t), x_n(t)\}. \quad (26)$$

where $l_n^C(t), l_n^D(t)$ represent the discrete actions of choosing communication and deception bands, denoted as $a_n^{dis}(t) = \{l_n^C(t), l_n^D(t)\}$. $x_n(t)$ represents the continuous action of allocating communication power, represented as $a_n^{con}(t) = x_n(t)$.

- 4) *Reward* \mathcal{R} : In the context of (10) and the global state, the reward should be designed to maximize SINR while avoiding the co-channel interference and the wastage of frequency bands. Since enforcing hard constraints (10c)-(10e) directly in decentralized MARL is challenging (each LU cannot instantly know other LUs' band selections due to the absence of real-time coordination), we reformulate them as soft penalties in the reward function to guide the policy toward constraint satisfaction. This design allows the distributed agents to learn cooperative behavior that gradually converges to the desired constraints through trial and error. Thus, the reward is defined as:

$$\mathcal{R} = w_1 \sum_{n=1}^N \text{SINR}_n(t) + w_2(k^C(t) - N) + w_3(1 - k^D(t)), \quad (27)$$

where w_i represents the weight for the i -th element. The definition of (27) has a triple physical meanings. First, it aims to maximize the received SINR at the UAV end, corresponding to the first part of (27). Second, it aims to mitigate the interference caused by multiple LUs utilizing the same communication band, corresponding to the second part of (27) which penalizes when the number of unique communication bands, $k^C(t)$, is less than N . Finally, it encourages all LUs to utilize a single deception channel to prevent spectrum wastage, corresponding to the third part of (27) which penalizes when the number of deception bands $k^D(t)$ is greater than 1.

C. Detailed Algorithm of CMAJD

However, in real UAV-assisted communication scenarios, global information is unavailable and the computational capabilities of UAVs are limited. To address the jamming deception problem formulated as an MDP, we design a MARL-based

method under the CTDE framework. The core idea is to deliberately expose a deception frequency band, attracting part of the MU's jamming power and thereby mitigating its adverse impact on legitimate communications. Under the CTDE framework, the UAV performs centralized performance evaluation while LUs execute distributed policies based on local observations. This approach avoids high-dimensional action decision-making at the UAV and enables efficient, low-dimensional cooperative jamming deception.

The CTDE framework, illustrated in Fig. 2, employs a centralized value network on the UAV and distributed policy networks on each LU. In each time slot, each LU selects an action according to its local policy:

$$a_n(t) \sim \pi(\cdot | o_n(t); \xi_n), \quad (28)$$

where ξ_n denotes the parameters of the policy network. The UAV then aggregates all LU actions into a global state:

$$\mathcal{S}(t) \leftarrow a_1(t), a_2(t), \dots, a_N(t), \quad (29)$$

and evaluates the anti-jamming performance using the value network $v(\mathcal{S}(t); \zeta)$, parameterized by ζ . During training, both the value and policy networks are updated collaboratively. During execution, each LU acts independently using its local policy network, avoiding centralized action aggregation. The benefits of this architecture are twofold: 1) centralized training applies a unified value function across all LUs, preventing spectrum competition and overcoming limitations of single-agent reinforcement learning; 2) distributed execution significantly reduces computational complexity. Both policy and value networks are composed of fully-connected layers with ReLU activations, orthogonal weight initialization, and layer normalization. Optionally, recurrent layers may be incorporated for temporal feature extraction.

Then, as shown in Fig. 2, based on the CTDE architecture, we design the CMAJD method to deceive jamming. This method employs the Proximal Policy Optimization (PPO) algorithm, an actor-critic method, for stable policy updates [3]. The advantage estimation follows an advantage actor-critic (A2C) style, while the policy update uses PPO's clipping objective. The detailed updated process of CMAJD is shown in Algorithm 1. At the beginning, the value network parameter set ζ and the policy network parameter set ξ_n of each LU are randomly initialized. Besides, the target value network is also initialized with random parameter set ζ' . The above sets define the weights and biases of each neuron in the neural network. In the specific update process, each LU observes local state $o_n(t)$, selects bands and allocates power for communication and deception based on the hybrid policy network:

$$a_n^{dis}(t) \sim \pi^{dis}(\cdot | o_n(t); \xi_n^{dis}), \quad (30)$$

$$a_n^{con}(t) \sim \pi^{con}(\cdot | o_n(t); \xi_n^{con}). \quad (31)$$

After executing the action $a_n(t)$, each LU can obtain the local observation of next timeslot $o_n(t+1)$. Then, the UAV will collect all the current and future local observations from all LUs, compiling them into the current global state $\mathcal{S}(t)$ and the next global state $\mathcal{S}(t+1)$, respectively. Then, the values of current and next states can be evaluated with the value

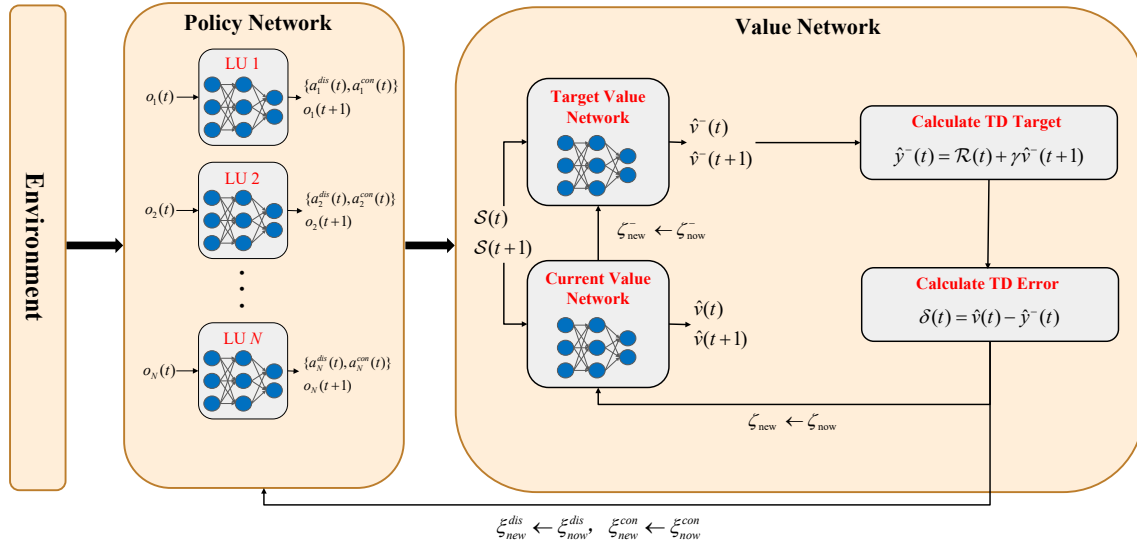


Fig. 2. The illustration of CMAJD proposed in this paper.

network, which can be shown as:

$$\hat{v}(t) = v(\mathcal{S}(t); \zeta_{\text{now}}), \quad (32)$$

$$\hat{v}^-(t+1) = v(\mathcal{S}(t+1); \zeta_{\text{now}}^-), \quad (33)$$

where $v(\mathcal{S}(t+1); \zeta_{\text{now}}^-)$ is the target value network, guiding the update of the current value network $v(\mathcal{S}(t+1); \zeta_{\text{now}})$. The update process initially involves calculating the target value for the temporal difference (TD) from the estimated value of $v(\mathcal{S}(t+1); \zeta_{\text{now}}^-)$:

$$\hat{y}^-(t) = \mathcal{R}(t) + \gamma \hat{v}^-(t+1), \quad (34)$$

where $\gamma \in (0, 1)$ denotes the discount factor. Based on this, we can compute the TD-error between the targeted value and the TD target value, referred to as the advantage function:

$$\delta(t) = \hat{v}(t) - \hat{y}^-(t). \quad (35)$$

The advantage function can be utilized helps to enhance the efficiency of network updates and reduce overfitting caused by excessive variance. Based on the advantage function, parameter set of the current value network can be updated as:

$$\zeta_{\text{new}}^{\text{con}} = \zeta_{\text{now}}^{\text{con}} - \lambda^C \cdot \delta(t) \cdot \nabla_{\zeta_{\text{now}}^{\text{con}}} \hat{v}(t), \quad (36)$$

where λ^C represents the learning rate of the current value network, and $\nabla_{\zeta_{\text{now}}^{\text{con}}} \hat{v}(t)$ means the gradient of $\hat{v}(t)$ with respect to $\zeta_{\text{now}}^{\text{con}}$. Subsequently, the parameter set of target value network can be updated as:

$$\zeta_{\text{new}}^- = \tau \cdot \zeta_{\text{new}}^{\text{con}} + (1 - \tau) \cdot \zeta_{\text{now}}^-, \quad (37)$$

where τ is a positive real number between (0, 1). By integrating (34), (35), (36) and (37), parameters of both current value network and target value network can be updated. Finally, based on the advantage function, the policy network of each LU can be updated with the proximal policy optimization

algorithm, which is shown as:

$$\zeta_{\text{new}}^{\text{dis}} = \arg \max_{\zeta_{\text{dis}}} \min \left(\frac{\pi(o_n(t); \zeta_{\text{new}}^{\text{dis}})}{\pi(o_n(t); \zeta_{\text{now}}^{\text{dis}})} \delta(t), \text{clip} \left(\frac{\pi(o_n(t); \zeta_{\text{new}}^{\text{dis}})}{\pi(o_n(t); \zeta_{\text{now}}^{\text{dis}})}, 1 - \varepsilon, 1 + \varepsilon \right) \delta(t) \right), \quad (38)$$

$$\zeta_{\text{new}}^{\text{con}} = \arg \max_{\zeta_{\text{con}}} \min \left(\frac{\pi(o_n(t); \zeta_{\text{new}}^{\text{con}})}{\pi(o_n(t); \zeta_{\text{now}}^{\text{con}})} \delta(t), \text{clip} \left(\frac{\pi(o_n(t); \zeta_{\text{new}}^{\text{con}})}{\pi(o_n(t); \zeta_{\text{now}}^{\text{con}})}, 1 - \varepsilon, 1 + \varepsilon \right) \delta(t) \right), \quad (39)$$

where ε is a small positive number. In the PPO algorithm, the clipping ratio $\frac{\pi(o_n(t); \zeta_{\text{new}})}{\pi(o_n(t); \zeta_{\text{now}})}$ controls the degree of parameter update. When this ratio is either too large or too small, the update degree is constrained within the range of $1 - \varepsilon$ to $1 + \varepsilon$ to prevent overly large update. The process discussed above is repeated until the algorithm converges or reaches the iteration limit.

The proposed architecture, where the UAV hosts the centralized critic and each LU holds a decentralized actor, offers distinct advantages tailored for UAV-assisted anti-jamming scenarios:

- *Computational efficiency for the UAV:* By decomposing the high-dimensional global action space into local low-dimensional decisions, the UAV is relieved from the heavy burden of directly computing or searching over the joint action space. This is crucial for resource-constrained UAV platforms.
- *Communication robustness under jamming:* Each LU acts based solely on its local observations, eliminating any requirement for real-time, reliable information exchange among LUs—a channel particularly vulnerable under intentional jamming.
- *Global coordination through centralized guidance:* While execution is decentralized, the centralized critic at the UAV evaluates the joint effect of all LU actions and guides their policy updates. This ensures that the inde-

Algorithm 1: Multi-Agent Collaborative Jamming deception Algorithm

Require:

Channel parameters, UAV location, LU locations, MU locations.

Ensure:

Value network parameter set ζ , policy network parameter set ξ_n of each LU.

- 1: Randomly initialize the current and target value network parameters with ζ and ζ^- , randomly initialize the policy network parameters ξ_n for each LU, randomly initialize actions.
 - 2: **for** Episode = 1, 2, \dots , E **do**
 - 3: **for** $t = 1, T$ **do**
 - 4: Each LU interacts with the environment, obtains local observation $o_n(t)$, and then selects its action $a_n(t)$ based on (30) and (31).
 - 5: LUs execute actions, obtain the next observation $o_n(t+1)$.
 - 6: UAV collects local observations $\mathcal{O}(t)$ and $\mathcal{O}(t+1)$ from each user to obtain the global states $\mathcal{S}(t)$ and $\mathcal{S}(t+1)$.
 - 7: UAV calculates the current state value $\hat{v}(t)$ and the future state value $\hat{v}^-(t+1)$ based on (32) and (33).
 - 8: Based on $\mathcal{S}(t)$ and $\mathcal{S}(t+1)$, calculate the temporal difference target value $\hat{y}^-(t)$ using (34), and calculate the advantage function $\delta(t)$ using (35).
 - 9: UAV updates the current evaluation network using (36), and updates the target evaluation network using (37).
 - 10: Each LU updates its policy network parameters using (38) and (39).
 - 11: **end for**
 - 12: **end for**
 - 13: **return** $v(\mathcal{S}(t); \zeta), \pi(\cdot | o_n(t); \xi_n)$.
-

pendently learned policies converge toward a globally cooperative optimum, avoiding detrimental conflicts such as spectrum collisions.

Together, these features enable the CMAJD framework to achieve effective collaborative anti-jamming in a scalable, robust, and computationally feasible manner.

D. Computational Complexity

Unlike traditional centralized reinforcement learning approaches, CMAJD significantly reduces the decision-making complexity by decomposing the global action space into per-agent local decisions. To quantify this advantage, we compare with representative methods like [41], whose action space dimension grows exponentially as $O\left(\frac{L!}{(L-(N+1))!} \times \chi^N\right)$ (where L , N , and $\chi + 1$ denote the numbers of bands, users, and power discretization levels, respectively). In contrast, by distributing the decision-making, CMAJD reduces the per-agent decision dimension to $O\left(\frac{L!}{(L-2)!} + 1\right) \approx O(L^2)$, achieving a crucial exponential-to-polynomial reduction. This directly translates to the lower computational overhead and

enables real-time operation on resource-constrained UAVs, as supported by the empirical training time comparisons in Table IV of Section IV.

For example, the method in [25] used reinforcement learning to choose actions of all LUs and discretized the continuous power to $\chi + 1$ levels. Thus, the dimension of the action space of the method in [25] is $\frac{L!}{(L-(N+1))!} \times \chi^N$. We deconstruct global action into local actions and the dimension of each policy network is $\frac{L!}{(L-2)!} + 1$. In other words, the action space dimension of all LUs is reduced from $\frac{L!}{(L-(N+1))!} \times \chi^N$ to $N \times \left(\frac{L!}{(L-2)!} + 1\right)$. The main computational complexity of Algorithm 1 arises from the updating of neural network parameters. In line 9, the complexity of updating the parameters of the current and target value networks is $O(D_S \times h_{in} + h_{in} \times h_{out} + h_{out})$, where D_S represents the dimensionality of the state space $\mathcal{S}(t)$, and h_{in} and h_{out} represent the numbers of input and output nodes of the neural network, respectively. In line 10, the computational complexity of updating the discrete and continuous action networks using the neural network is $O(N \times (D_a \times h_{in} + h_{in} \times h_{out} + h_{out}))$, where D_a represents the dimensionality of discrete and continuous actions. Therefore, the computational complexity of Algorithm 1 is $O(E \times T \times (N + D_S \times h_{in} + h_{in} \times h_{out} + N \times D_a \times h_{in}))$.

IV. NUMERICAL RESULTS

In this section, we provide numerical results to evaluate the performance of the proposed CMAJD algorithm. The simulation scenario is a complex urban environment. The LUs and the MU are uniformly distributed within a circle of radius 100 m, with their coordinates (x, y) independently and uniformly sampled. The MU's mobility is modeled as a random walk: in each time slot, it moves with a uniformly distributed direction and a step length uniformly distributed in $[0, 1]$ m. To characterize the urban propagation, the environment-dependent parameters in the LoS probability model (2) are set to $a = 11.95$ and $b = 0.14$, following the standard urban model [43]. The complete set of simulation parameters is provided in Table III.

The reward weights w_1 , w_2 , and w_3 are critical to balancing the primary objective of SINR maximization against the soft constraints (avoiding communication band collisions and enforcing a single deception band). The weight w_1 for the SINR term is set to 1.0 as a baseline to prioritize communication quality, while the penalty weights w_2 and w_3 are determined through systematic grid search over the ranges $w_2 \in [0.5, 10]$ and $w_3 \in [1, 20]$, with the goal of achieving both high SINR and constraint satisfaction. The final values $w_2 = 2.0$ and $w_3 = 4.0$ are selected because they provide a sufficient penalty for constraint violations (as evidenced by the converged policies in Fig. 4, where all LUs choose distinct communication bands and a single deception band) while still allowing the SINR term to effectively guide the learning toward high-performance communication. We observe that setting w_2 and w_3 too low leads to frequent collisions and multiple deception bands, whereas excessively high weights cause the agents to prioritize constraint satisfaction at the expense of SINR. The

TABLE III
SIMULATION PARAMETERS

Parameter	Physical Meaning	Value
c	Speed of electromagnetic wave propagation	3×10^8 m/s
η_{LoS}	Large-scale attenuation factor for LoS links	0.5
η_{NLoS}	Large-scale attenuation factor for NLoS links	0.005
α_{g2a}	Ground-to-air path loss parameter	2
α_{g2g}	Ground-to-ground path loss parameter	2.8
E	Number of training cycles	500
T	Total number of time slots	100 s
P_{max}	Maximum transmit power of the user	10 dBw
τ	Value network update parameter	0.9
ϵ	PPO clipping parameter	0.1
σ^2	Power of the additive white Gaussian noise	-110 dBm

chosen values strike a favorable balance, as further validated by the near-optimal SINR results in Figs. 5–7.

Based on the model in [43], the optimal UAV altitude for maximizing the channel gain to cell-edge users is derived as $h_{opt} = 100 \times \tan \theta_{opt}$, where θ_{opt} is the elevation angle maximizing the gain. This yields $h_{opt} \approx 82.53$ m [2]. The MU's trajectory is defined to start at a random boundary point and reach another random boundary point after 100 time slots. Consequently, the channel responses $\bar{g}_n^{g2g}(t)$, $\bar{g}_n^{l2g}(t)$, and $\bar{g}_j^{g2a}(t)$ vary dynamically across time slots.

First, the convergence of Algorithm 1 is analyzed, as shown in Fig. 3. It displays the convergence of the CMAJD algorithm compared to SARL under different numbers of LUs [41], where the vertical axis represents the average reward calculated as the mean of all rewards over 100 time slots. SARL is a model-free reinforcement learning algorithm that employs an 8-layer fully connected neural network with ReLU activation functions. The algorithm generates its training data from game-theoretic optimal solutions. It can be observed that CMAJD exhibits good convergence in all simulation settings (converging to a stable value around 70 iterations and remaining stable in subsequent iterations). As the number of the LUs increases, the total power available for deception increases, thereby allowing the MU to waste more power in the deceptive channels, which in turn leads to a significant increase of the average rewards. Meanwhile, the SARL algorithm fails to converge to a feasible solution within the designed number of iterations due to the high dimensionality of the decision space. The adoption of the CTDE architecture significantly reduces the decision dimensionality, allowing the CMAJD method to converge quickly. Additionally, increasing the number of users does not cause an explosion in the dimensionality of the action space, thus the convergence speed does not decrease as the number of users increases. In summary, the curves in Fig. 3 demonstrate that the algorithm proposed in this paper has good convergence properties and can quickly converge to a stable and feasible solution.

Next, we analyze the effectiveness of the proposed method for the selection of LU communication bands and deception bands. As shown in Fig. 4, the selection strategies of commu-

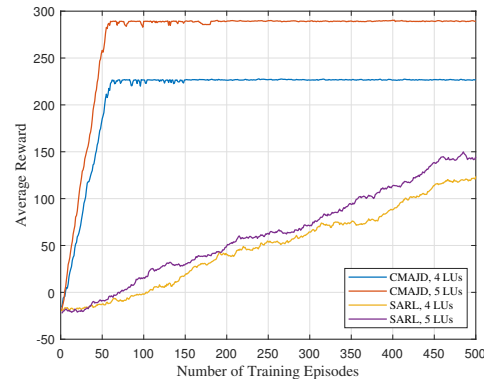
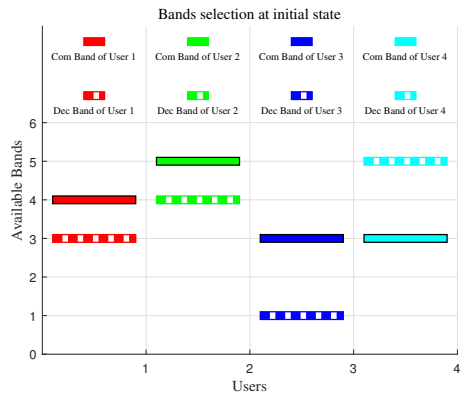


Fig. 3. The convergence of the proposed algorithm.

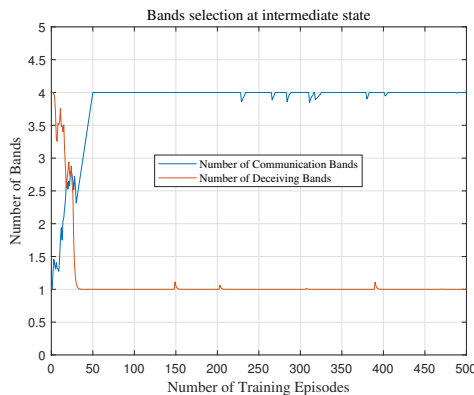
nication and deception bands for all LUs are displayed. From Fig. 4(a), it can be seen that initially, LUs randomly select the communication band and the spoofing band, respectively, which leads to conflicts between the communication bands among users or conflicts between the communication bands and the deception bands. There also exist inconsistencies in deception bands, resulting in wasted spectrum resources. As shown in Fig. 4(c), through training, the LUs can choose different communication bands and the same deception band that do not conflict with any LUs communication bands, thereby avoiding interference among all LUs while saving available bands. Moreover, Fig. 4(b) shows the changes in the numbers of communication and deception bands per iteration. It can be observed that initially, there may be conflicts in communication bands and wastage of deception bands, but as the number of iterations increases, the number of communication bands quickly rises to match the number of LUs, avoiding interference among users. Simultaneously, the number of deception bands quickly reduces to 1, indicating that all LUs use the same band to deceive MU, avoiding unnecessary wastage of the spectrum. This behavior is reinforced by the reward function, which includes penalty terms for deviations of both the number of communication bands and the number of deception bands from their ideal values.

The convergence analysis in Fig. 3 reveals that SARL, a single-agent reinforcement learning method, fails to converge in multi-user scenarios due to the exponential growth of the action space. Hence, SARL is not a viable candidate for performance comparison in our setting. Instead, we adopt two multi-agent reinforcement learning baselines that are more relevant to cooperative anti-jamming: independent multi-agent reinforcement learning (IMARL) and multi-agent proximal policy optimization (MAPPO). These methods enable us to assess the advantages of our CTDE-based CMAJD framework. Specifically, the following three benchmark schemes are designed to evaluate the performance of UAV communication systems:

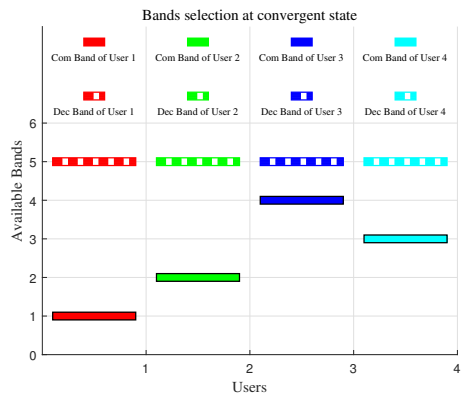
- 1) *SINR at the equilibrium*: Assuming that all environmental information is known, the optimal strategies for LU and MU can be obtained using the method introduced in Section III-A, and the SINR of the UAV communication system at this point is considered as the performance upper bound.



(a) Initial State



(b) Intermediate State



(c) Convergent State

Fig. 4. Band selection of all LUs at the initial state, the intermediate state and the convergent state.

- 2) *SINR of the independent multi-agent reinforcement learning (IMARL) method:* Unlike the scheme proposed in this paper, this scheme does not adopt a centralized training architecture. All LUs operate independently, evaluating their own value networks and policy networks. This approach suffers from spectrum competition and lack of coordination, often leading to suboptimal performance.
- 3) *SINR of the multi-agent proximal policy optimization*

(MAPPO) reinforcement learning method: This scheme follows the design described in [41], where the continuous action of power distribution is discretized into 10 levels, with only a discrete action network deployed at the LU end to select actions. The discretization leads to performance loss compared to continuous power allocation.

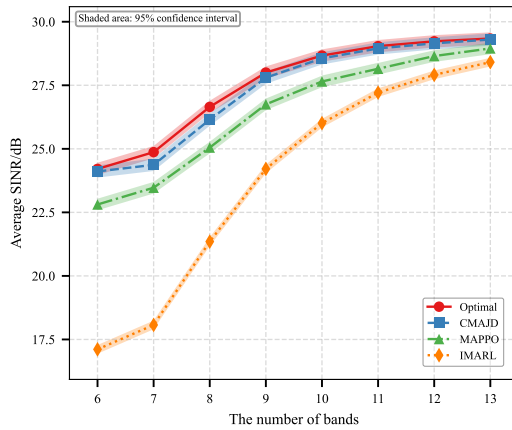
Table IV presents a comprehensive comparison of the convergence behavior and final performance of CMAJD and two representative baselines. As shown in the table, IMARL exhibits the slowest convergence and highest total training time (203 episodes, 694.26 s) due to its lack of centralized coordination, which leads to spectrum competition among independently acting LUs. MAPPO improves upon IMARL through its CTDE framework, converging in 114 episodes with 369.36 s total time, but suffers from performance degradation caused by its discretized power allocation. In contrast, CMAJD achieves the fastest convergence (65 episodes to 95%), the lowest total training time (206.05 s), and the highest performance metrics (24.37 dB SINR, 32.41 bit/s/Hz sum rate). Although CMAJD requires maintaining multiple networks, its average time per episode (3.17 s) is comparable to or slightly lower than both baselines. These results demonstrate that our approach with continuous power allocation and coordinated deception strategy effectively balances computational efficiency and anti-jamming performance, making it particularly suitable for resource-constrained UAV platforms.

Fig. 5 compares both the average SINR and the sum rate (in bit/s/Hz) at the UAV when there are 4 LUs under different numbers of available frequency bands, with noise power at -110 dBm. Each data point and its corresponding confidence interval in Fig. 5 are calculated from 100 independent simulation runs with random initializations. Fig. 5(a) shows the average SINR when the jamming power $P_{\max}^J = 100$ W. It can be seen that the SINR increases with the number of available bands because more band choices provide a better opportunity to find optimal allocations that mitigate interference. Similarly, Fig. 5(b) shows the sum rate, which follows a similar increasing trend as SINR, confirming that higher SINR directly translates to improved throughput. The marginal gain in both metrics diminishes when the number of available bands exceeds a critical point, as the system's ability to exploit band diversity becomes saturated. Importantly, the confidence intervals show that CMAJD's performance is statistically comparable to the theoretical upper bound (their confidence intervals overlap), while clearly outperforming both MAPPO and IMARL (no overlap in confidence intervals).

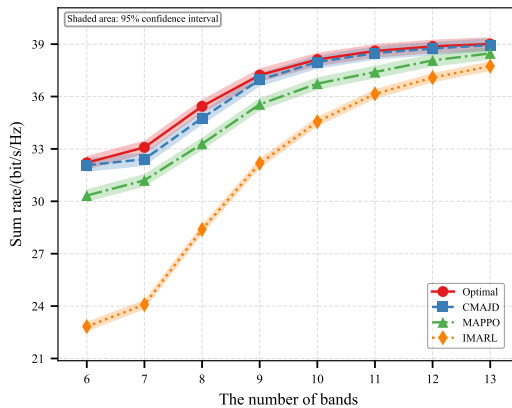
Fig. 6 shows the performance comparison when there are 7 frequency bands available and the number of LUs varies. Again, the results are averaged over 100 independent runs with 95% confidence intervals. Fig. 6(a) presents the average SINR with $P_{\max}^J = 100$ W. As the number of LUs increases, the SINR improves significantly because more users contribute deception power, thereby attracting a larger portion of the MU's jamming power away from the communication bands. Fig. 6(b) shows the corresponding sum rate, confirming that the collaborative deception strategy effectively converts increased user participation into higher system throughput.

TABLE IV
COMPUTATIONAL EFFICIENCY AND PERFORMANCE COMPARISON ($N = 4, L = 7$)

Method	Avg. Time per Ep. (s)	Eps. to 95% Conv.	Total Time (s)	Avg. SINR (dB)	Sum Rate (bit/s/Hz)
IMARL	3.42	203	694.26	18.07	24.08
MAPPO [41]	3.24	114	369.36	23.47	31.2
CMAJD (Ours)	3.17	65	206.05	24.37	32.41



(a) Average SINR vs. number of frequency bands ($P_{\max}^J = 100$ W, $N = 4$).

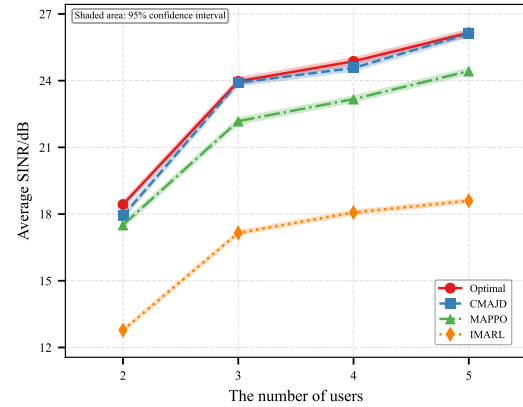


(b) Sum rate vs. number of frequency bands ($P_{\max}^J = 100$ W, $N = 4$).

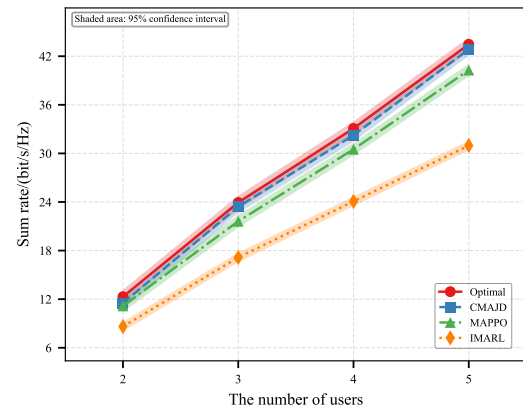
Fig. 5. Performance comparison at the UAV end with different numbers of available frequency bands. The results are averaged over 100 independent simulation runs; the shaded regions represent the 95% confidence intervals.

The confidence intervals demonstrate that CMAJD achieves performance statistically similar to the optimal bound (overlapping confidence intervals), while significantly outperforming IMARL and MAPPO (non-overlapping confidence intervals).

To further investigate the robustness of the proposed method against increasing jamming threat, Fig. 7 presents the system performance as the MU's maximum jamming power P_{\max}^J varies from 100 W to 300 W, with $N = 4$ LUs and $L = 7$ available bands. Fig. 7(a) shows that the average SINR decreases as P_{\max}^J increases, which is expected because an



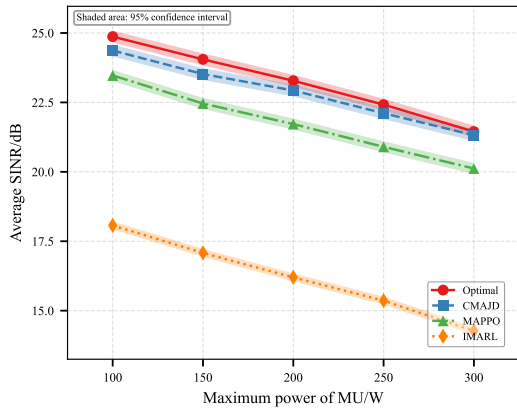
(a) Average SINR vs. number of LUs ($P_{\max}^J = 100$ W, $L = 7$).



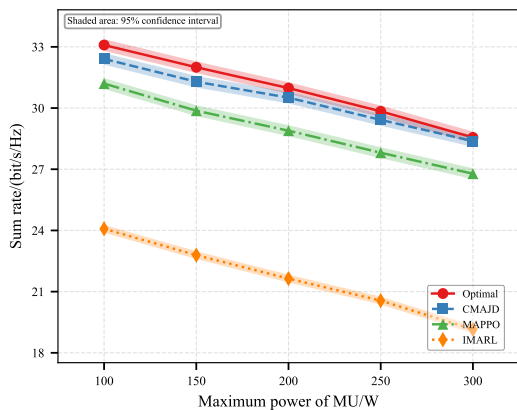
(b) Sum rate vs. number of LUs ($P_{\max}^J = 100$ W, $L = 7$).

Fig. 6. Performance comparison at the UAV end with different numbers of LUs. The results are averaged over 100 independent simulation runs; the shaded regions represent the 95% confidence intervals.

increased jamming power raises the interference level at the UAV. However, CMAJD demonstrates a graceful degradation and maintains a performance level close to the theoretical optimum (their confidence intervals overlap). Similarly, Fig. 7(b) illustrates the corresponding sum rate, which follows the same declining trend but again confirms that CMAJD consistently delivers higher throughput than IMARL and MAPPO, with non-overlapping confidence intervals indicating statistically significant superiority. These results underscore the adaptability of CMAJD to varying jamming conditions and its effectiveness in preserving communication quality even when the MU's operational range (reflected in its available



(a) Average SINR vs. MU's maximum jamming power P_{\max}^J ($N = 4$, $L = 7$, $P_{\max}^J = 100$ W).



(b) Sum rate vs. MU's maximum jamming power P_{\max}^J ($N = 4$, $L = 7$).

Fig. 7. Performance comparison at the UAV end under varying MU jamming power. The results are averaged over 100 independent simulation runs; the shaded regions represent the 95% confidence intervals.

power) expands.

Combining the aforementioned numerical results, it can be concluded that the proposed CMAJD method can rapidly converge to a feasible solution. It achieves jamming deception using a single frequency band while ensuring the normal transmission of the UAV-assisted wireless communication system. Additionally, it can approach the theoretical upper bound of both SINR and sum rate for anti-jamming, thus ensuring the communication performance of the system under various conditions.

V. CONCLUSION

In response to the challenge that UAVs with limited payload capacities may face the high complexity of anti-jamming decision-making, this paper proposes an MARL jamming deception method based on the CTDE architecture. This approach targets reactive jamming by setting actively exposed deception frequency band to attract part of the jamming power. First, the anti-jamming process is modeled as a Stackelberg

game. Then, optimal strategies for LUs and MU are determined under the assumption that all environmental information is known. Finally, we focus on the real-world scenario where the information is unknown and designed the CMAJD method. Specifically, distributed policy networks are deployed at LUs. LUs make decisions based on their own observations according to the policy networks. This approach achieves a dimension reduction of high-dimensional global anti-jamming decisions. At the UAV end, a centralized value network is deployed to evaluate the impact of distributed decisions on global performance, guiding the update of policy networks. Simulation results shows that the CMAJD method can converge quickly. Meanwhile, different communication bands can be effectively selected to avoid interference, and the same deception band is selected to save spectrum resources. Moreover, the anti-jamming performance of CMAJD can approach the performance upper bound, significantly outperforming other benchmark schemes.

REFERENCES

- [1] Z. Xiao, P. Xia, and X.-G. Xia, "Enabling UAV cellular with millimeter-wave communication: Potentials and approaches," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 66–73, May 2016.
- [2] C. Zhang, L. Zhang, L. Zhu, T. Zhang, Z. Xiao, and X.-G. Xia, "3D deployment of multiple UAV-mounted base stations for UAV communications," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2473–2488, Apr. 2021.
- [3] Z. Xia, J. Du, J. Wang, C. Jiang, Y. Ren, G. Li, and Z. Han, "Multi-agent reinforcement learning aided intelligent UAV swarm for target tracking," *IEEE Trans. Veh. Technol.*, vol. 71, no. 1, pp. 931–945, Jan. 2022.
- [4] Z. Han, N. Marina, M. Debbah, and A. Hjørungnes, "Physical layer security game: Interaction between source, eavesdropper, and friendly jammer," *EURASIP J. Wireless Commun. Netw.*, vol. 2009, pp. 1–10, Aug. 2009.
- [5] W. Xu, W. Trappe, Y. Zhang, and T. Wood, "The feasibility of launching and detecting jamming attacks in wireless networks," in *Proc. ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, Urbana-Champaign, IL, 2005.
- [6] A. Toma, A. Krayani, M. Farrukh, H. Qi, L. Marcenaro, Y. Gao, and C. S. Regazzoni, "AI-based abnormality detection at the PHY-layer of cognitive radio by learning generative models," *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 1, pp. 21–34, Mar. 2020.
- [7] V. Hassija, V. Chamola, A. Agrawal, A. Goyal, N. C. Luong, D. Niyato, F. R. Yu, and M. Guizani, "Fast, reliable, and secure drone communication: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 4, pp. 2802–2832, 4th Quart. 2021.
- [8] X. Lu, L. Xiao, P. Li, X. Ji, C. Xu, S. Yu, and W. Zhuang, "Reinforcement learning-based physical cross-layer security and privacy in 6G," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 425–466, 1st Quart. 2023.
- [9] A. Fotouhi, H. Qiang, M. Ding, M. Hassan, L. G. Giordano, A. Garcia-Rodriguez, and J. Yuan, "Survey on UAV cellular communications: Practical aspects, standardization advancements, regulation, and security challenges," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3417–3442, 4th Quart. 2019.
- [10] Y. Zhi, Z. Fu, X. Sun, and J. Yu, "Security and privacy issues of UAV: A survey," *Mobile Netw. Appl.*, vol. 25, pp. 95–101, Feb. 2020.
- [11] X. Tan, S. Su, and X. Sun, "Research on narrowband interference suppression technology of UAV network based on spread spectrum communication," in *Proc. IEEE Int. Conf. Artif. Intell. Inf. Syst. (ICAIS)*, Haikou, China, Mar. 2020, pp. 335–338.
- [12] B. M. Todorovic and V. D. Orlic, "Direct sequence spread spectrum scheme for an unmanned aerial vehicle PPM control signal protection," *IEEE Commun. Lett.*, vol. 13, no. 10, pp. 727–729, Oct. 2009.
- [13] A. Masmoudi, F. Bellili, S. Affes, and A. Ghayeb, "Maximum likelihood time delay estimation from single- and multi-carrier DSSS multipath MIMO transmissions for future 5G networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 4851–4865, Aug. 2017.
- [14] A. Tayebi, S. Berber, and A. Swain, "Security enhancement of fix chaotic-DSSS in WSNs," *IEEE Commun. Lett.*, vol. 22, no. 4, pp. 816–819, Apr. 2018.

- [15] Y. Dong, C. He, Z. Wang, and L. Zhang, "Radio map assisted path planning for UAV anti-jamming communications," *IEEE Signal Process. Lett.*, vol. 29, pp. 607–611, Feb. 2022.
- [16] J. Bao and L. Ji, "Frequency hopping sequences with optimal partial hamming correlation," *IEEE Trans. Inf. Theory*, vol. 62, no. 6, pp. 3768–3783, Jun. 2016.
- [17] G. Reus-Muns, M. Diddi, C. Singhal, H. Singh, and K. R. Chowdhury, "Flying among stars: Jamming-resilient channel selection for UAVs through aerial constellations," *IEEE Trans. Mobile Comput.*, vol. 22, no. 3, pp. 1246–1262, Mar. 2023.
- [18] M. Z. Win, "A unified spectral analysis of generalized time-hopping spread-spectrum signals in the presence of timing jitter," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 9, pp. 1664–1676, Dec. 2002.
- [19] J. Yu, Y. Gong, J. Fang, R. Zhang, and J. An, "Let us work together: Cooperative beamforming for UAV anti-jamming in space-air-ground networks," *IEEE Internet Things J.*, vol. 9, no. 17, pp. 15607–15617, Sep. 2022.
- [20] H. Pirayesh and H. Zeng, "Jamming attacks and anti-jamming strategies in wireless networks: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 2, pp. 767–809, 2nd Quart. 2022.
- [21] X. Lu, L. Xiao, C. Dai, and H. Dai, "UAV-aided cellular communications with deep reinforcement learning against jamming," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 48–53, Aug. 2020.
- [22] H. Zhang, C. Lu, H. Tang, X. Wei, L. Liang, L. Cheng, W. Ding, and Z. Han, "Mean-field-aided multiagent reinforcement learning for resource allocation in vehicular networks," *IEEE Internet Things J.*, vol. 10, no. 3, pp. 2667–2679, Feb. 2023.
- [23] X. Pei, X. Wang, J. Yao, C. Yao, J. Ge, L. Huang, and D. Liu, "Joint time-frequency anti-jamming communications: A reinforcement learning approach," in *Proc. IEEE Wireless Commun. Signal Process. (WCSP)*, Xi'an, China, Oct. 2019, pp. 1–6.
- [24] C. Li, S. Yang, and T. T. Nguyen, "A self-learning particle swarm optimizer for global optimization problems," *IEEE Trans. Syst., Man, Cybern. B.*, vol. 42, no. 3, pp. 627–646, Jun. 2012.
- [25] A. Pourranjbar, G. Kaddoum, A. Ferdowsi, and W. Saad, "Reinforcement learning for deceiving reactive jammers in wireless networks," *IEEE Trans. Commun.*, vol. 69, no. 6, pp. 3682–3697, Jun. 2021.
- [26] M. E. Mkiramweni, C. Yang, J. Li, and W. Zhang, "A survey of game theory in unmanned aerial vehicles communications," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3386–3416, 4th Quart. 2019.
- [27] Z. Han, D. Niyato, W. Saad, T. Başar, and A. Hjørungnes, *Game Theory in Wireless and Communication Networks: Theory, Models, and Applications*. Cambridge, U.K.: Cambridge University Press, 2012.
- [28] X. Tang, P. Ren, Y. Wang, Q. Du, and L. Sun, "Securing wireless transmission against reactive jamming: A stackelberg game framework," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, San Diego, CA, Dec. 2015, pp. 1–6.
- [29] D. Yang, G. Xue, J. Zhang, A. Richa, and X. Fang, "Coping with a smart jammer in wireless networks: A stackelberg game approach," *IEEE Trans. Wireless Commun.*, vol. 12, no. 8, pp. 4038–4047, Aug. 2013.
- [30] Y. Zhang, Y. Xu, Y. Xu, Y. Yang, Y. Luo, Q. Wu, and X. Liu, "A multi-leader one-follower stackelberg game approach for cooperative anti-jamming: No pains, no gains," *IEEE Commun. Lett.*, vol. 22, no. 8, pp. 1680–1683, Aug. 2018.
- [31] S. D'Oro, L. Galluccio, G. Morabito, S. Palazzo, L. Chen, and F. Martignon, "Defeating jamming with the power of silence: A game-theoretic analysis," *IEEE Trans. Wireless Commun.*, vol. 14, no. 5, pp. 2337–2352, May 2015.
- [32] Y. Gao, Y. Xiao, M. Wu, M. Xiao, and J. Shao, "Game theory-based anti-jamming strategies for frequency hopping wireless communications," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5314–5326, Aug. 2018.
- [33] 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 J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1587–1599, May 2022.
- [34] Y. Shao, Y. Cai, T. Wang, Z. Guo, P. Liu, J. Luo, and D. Gündüz, "Learning-based autonomous channel access in the presence of hidden terminals," *IEEE Trans. Mobile Comput.*, vol. 23, no. 5, pp. 3680–3695, May 2024.
- [35] C. Huang, G. Chen, P. Xiao, Y. Xiao, Z. Han, and J. A. Chambers, "Joint offloading and resource allocation for hybrid cloud and edge computing in SAGINs: A decision assisted hybrid action space deep reinforcement learning approach," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 5, pp. 1029–1043, May 2024.
- [36] L. Zhao, M. Valero, S. Pouriyeh, F. Li, L. Guo, and Z. Han, "A decentralized communication-efficient federated analytics framework for connected vehicles," *IEEE Trans. Veh. Technol.*, vol. 73, no. 7, pp. 10856–10861, Jul. 2024.
- [37] Z. Yin, Y. Lin, Y. Zhang, Y. Qian, F. Shu, and J. Li, "Collaborative multiagent reinforcement learning aided resource allocation for UAV anti-jamming communication," *IEEE Internet Things J.*, vol. 9, no. 23, pp. 23995–24008, Dec. 2022.
- [38] L. Miuccio, S. Riolo, S. Samarakoon, M. Bennis, and D. Panno, "On learning generalized wireless MAC communication protocols via a feasible multi-agent reinforcement learning framework," *IEEE Trans. Mach. Learn. Commun. Netw.*, vol. 2, pp. 298–317, Feb. 2024.
- [39] J. Akram, A. Anaissi, R. S. Rathore, R. H. Jhaveri, and A. Akram, "Digital twin-driven trust management in Open RAN-based spatial crowdsourcing drone services," *IEEE Trans. Green Commun. Netw.*, vol. 8, no. 3, pp. 1061–1075, Sep. 2024.
- [40] S. Tariq, U. Khalid, B. E. Arfeto, T. Q. Duong, and H. Shin, "Integrating sustainable big AI: Quantum anonymous semantic broadcast," *IEEE Wireless Commun.*, vol. 31, no. 3, pp. 86–99, Jun. 2024.
- [41] A. Pourranjbar, G. Kaddoum, and K. Aghababaiyan, "Deceiving-based anti-jamming against single-tone and multitone reactive jammers," *IEEE Trans. Commun.*, vol. 70, no. 9, pp. 6133–6148, Sep. 2022.
- [42] Y. Zeng, Q. Wu, and R. Zhang, "Accessing from the sky: A tutorial on UAV communications for 5G and beyond," *Proc. IEEE*, vol. 107, no. 12, pp. 2327–2375, Dec. 2019.
- [43] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal lap altitude for maximum coverage," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 569–572, Dec. 2014.
- [44] M. K. Hanawal, D. N. Nguyen, and M. Krunz, "Cognitive networks with in-band full-duplex radios: Jamming attacks and countermeasures," *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 1, pp. 296–309, Mar. 2020.
- [45] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge University Press, 2004.