SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

January 2022, Vol. 65 112304:1–112304:13
https://doi.org/10.1007/s11432-020-3170-2

Reinforcement learning based energy efficient robot relay for unmanned aerial vehicles against smart jamming

Xiaozhen LU[1], Jingfang JIE[1], Zihan LIN[1], Liang XIAO[1*], Jin LI[2] & Yanyong ZHANG[3]

[1]*Department of Information and Communication Engineering, Xiamen University, Xiamen* 361005*, China;*
[2]*School of Computer Science and Educational Software, Guangzhou University, Guangzhou* 510006*, China;*
[3]*School of Computer Science and Technology, University of Science and Technology of China, Hefei* 230027*, China*

Received 30 August 2020/Revised 26 November 2020/Accepted 4 January 2021/Published online 27 December 2021

Abstract Unmanned aerial vehicles (UAVs) with limited energy resources, severe path loss, and shadowing to the ground base stations are vulnerable to smart jammers that aim to degrade the UAV communication performance and exhaust the UAV energy. The UAV anti-jamming communication performance, such as the outage probability, degrades if the robot relay is not aware of the jamming policies and the UAV network topology. In this paper, we propose a robot relay scheme for UAVs against smart jamming, which combines reinforcement learning with a function approximation approach named tile coding, to jointly optimize the robot moving distance and relay power with the unknown jamming channel states and locations. The robot mobility and relay policy are chosen based on the received jamming power, the robot received signal quality, location and energy consumption, and the bit error rate of the UAV messages. We also present a deep reinforcement learning version for the robot with sufficient computing resources. It uses three deep neural networks to choose the robot mobility and relay policy with reduced sample complexity, so as to avoid exploring dangerous policies that lead to the high outage probability of the UAV messages. The network architecture of the three networks is designed with fully connected layers instead of convolutional layers to reduce the computational complexity, which is analyzed by theoretical analyses. We provide the performance bound of the proposed schemes in terms of the bit error rate, robot energy consumption and utility based on a game-theoretic study. Simulation results show that the performance of our proposed relay schemes, including the bit error rate, the outage probability, and the robot energy consumption outperforms the existing schemes.

Keywords unmanned aerial vehicles, relay, jamming, game theory, reinforcement learning

Citation Lu X Z, Jie J F, Lin Z H, et al. Reinforcement learning based energy efficient robot relay for unmanned aerial vehicles against smart jamming. Sci China Inf Sci, 2022, 65(1): 112304, https://doi.org/10.1007/s11432-020-3170-2

1 Introduction

Unmanned aerial vehicle (UAV) to ground communications are more vulnerable to jamming attacks than terrestrial communications [1], owing to the limited UAV energy resources, severe path loss, and shadowing caused by high altitude and mobility [2, 3]. Jammers can send jamming signals that contain faked commands or information to prevent the base station (BS) from receiving the UAV messages. Particularly, a smart jammer can even use universal software radio peripherals to continuously eavesdrop on the status of channels, and apply reinforcement learning (RL) algorithms to choose its jamming strategy accordingly [4]. Compared with the static jammers and random jammers, smart jammers are more reactive and aim to interrupt the transmissions from UAVs to BSs, exhaust UAVs' battery, and further launch the denial of service attacks.

Typical anti-jamming techniques for wireless communications, such as frequency hopping, are not applicable to UAVs with limited energy and bandwidth resources [2, 5]. Owing to the high mobility,

* Corresponding author (email: lxiao@xmu.edu.cn)

UAVs can choose their trajectories to resist jamming. For instance, a smart UAV trajectory design scheme (STUD) in [6] optimizes the next waypoint of the UAV to improve the service quality of the sensing task given the required waypoints and save the energy consumption. However, this scheme suffers from severe path loss and requires the UAV to move within a specific area, which may lead to sensing task failure under strong jamming attacks.

Robots can be used as relays to help wireless networks improve the transmission quality against jamming owing to their faster deployment and flexible mobility [7]. Compared with the fixed relays such as the ground node with limited coverage that are more expensive to deploy [7] and the UAV relays that have specific tasks such as the sensing duty [8], the robots can help UAVs relay the messages by changing their locations or transmit power to improve the communication performance [9, 10]. For instance, the relay scheme based on spectral graph theory in [9] that optimizes the mobility strategy for less energy consumption assumes full knowledge of the jammer location and channel states, which are rarely known by the robot relay in a dynamic UAV-ground communication system under severe jamming.

Inspired by the efficiency in video and strategy board games, reinforcement learning has been used to optimize the transmit power and the trajectory for cellular networks [11], mobile communication systems [12], and UAV networks [13] without the prior knowledge of the network topology and channel states of jammers. For example, the RL-based trajectory control scheme (QTC) designed in [11] applies Q-learning to choose the moving strategy based on the other UAVs' location to reduce the packet loss rate. Nevertheless, this relay scheme cannot be directly applied in the robot aided UAV-ground communication system against jamming attacks and thus suffers from the performance degradation.

In this paper, we propose an RL-based robot relay scheme for UAVs against smart jamming. Instead of directly using a standard Q-learning such as QTC in [11], the robot relay scheme applies RL to jointly optimize the moving distance and relay power, and uses a function approximation approach named tile coding [14] to map all the quantized states into a number of tilings to reduce the storage overhead. More specifically, the robot mobility and relay policy are chosen based on the received jamming power, the robot received signal strength indicator (RSSI), location and energy consumption, and the bit error rate (BER) of the UAV messages, which are used to build the state. Similar to [15], a hash function is used to map the current state into several active tiles, in which each tile represents one of the transmission features of the state. This scheme aims to increase the robot utility that contains the signal-to-interference-plus-noise ratio (SINR) and the mobility and transmission energy consumption.

We present a deep RL version for the robot that has enough computing resources to further improve the anti-jamming performance. By combining safe reinforcement learning with deep learning, this scheme avoids choosing dangerous robot mobility and relay policies that cause a high outage probability of the UAV messages and uses three deep neural networks, i.e., an online network, a risk network and a target network, to extract the relay anti-jamming features. More specifically, this relay scheme uses the BER of the UAV messages and the quality of service threshold as the basis to evaluate the risk values of each mobility and relay policy under a specific state. All the three neural networks have the same architecture, including an input layer, a hidden layer, and an output layer. The online network that outputs the Q-values and the risk network that outputs the long-term risk values are used to choose the mobility and relay policy, and the target network is used to reduce the overestimation of the Q-values. According to [6], fully connected layers that provide lower computational and sample complexity than convolutional layers in [16] are more effective for wireless anti-jamming communications. Thus, this scheme uses fully connected layers in the three networks for the robot relay scenario to extract the anti-jamming features in the state and thus reduce the exploration overhead.

We formulate a robot relay anti-jamming game, in which the robot chooses its mobility and relay policy for lower BER, outage probability, and energy consumption, and the jammer selects its jamming power to degrade the robot utility with less jamming power. The lower bounds that contain the BER and energy consumption are derived based on the Nash equilibrium. Simulation results based on the UAV-ground transmission model in [17] and the log-distance path loss channel model in [18] show that the performance of the proposed schemes exceeds STUD in [6] and QTC in [11] by reducing the BER and the outage probability, saving the robot energy consumption, and increasing the robot utility.

The rest of the paper is organized as follows. We review the related work in Section 2, followed by the robot aided UAV-ground communication model in Section 3. We propose two robot relay schemes for the UAV against smart jamming in Sections 4 and 5. Performance analysis and the simulation results are provided in Sections 6 and 7, and the conclusion is drawn in Section 8.

2 Related work

Power allocation is important for wireless networks against the random jamming [19], the fixed jamming [20] and the smart jamming attacks [21]. For instance, the orthogonal frequency-division multiplexing based Internet of Things system in [19] formulates a Colonel Blotto game and applies an evolutionary algorithm to find the Nash equilibrium of the formulated game with a lower BER against a random jammer. The stochastic approximation based anti-jamming scheme in [20] jointly optimizes the power allocation strategies of the users with energy harvesting based on the channel vector of the jammer-user link for a higher transmission rate and a lower computational complexity. The Bayesian game based power control scheme in [21] applies the duality optimization theory to derive the Stackelberg equilibrium against a smart jammer.

Trajectory planning has been used to resist jamming for UAV networks. The UAV in [6] applies a Q-network that consists of three fully connected layers to optimize the trajectory for a lower BER and less energy consumption to resist a greedy based smart jammer that can observe the channel status. The user with a single antenna in [22] that applies deep recurrent Q-network to choose the moving trajectory with the incomplete channel state information for less energy consumption achieves the SINR upper bound against a smart jammer that supports deep learning.

Relay can help improve the communication performance of wireless networks. The dynamic adaptive anti-jamming scheme in [9] uses spectral graph theory to select the relay position based on the Cheeger-like inequalities to increase the achievable capacity of the cognitive radio network against a static jammer and a mobile jammer. The multiple-input single-output system in [23] uses the distributed pricing-based optimization to choose the channel access probability of each relay to resist a reactive jammer with limited energy that can eavesdrop on all the available channels simultaneously. The cooperative relay beamforming scheme in [24] applies the relaxation approximation method to choose the relay vehicles and the corresponding beamformer for a higher network capacity against the radio frequency jamming attacks. The secure UAV-aided non-orthogonal multiple access (NOMA) scheme as proposed in [25] applies convex optimization to optimize the transmit power and precoding vectors for higher network throughput and secrecy rate. The mmWave-enabled NOMA-UAV system as designed in [26] jointly optimizes the UAV placement, the hybrid precoding, and the transmit power of users to further improve the energy efficiency and sum rate.

Reinforcement learning helps wireless networks resist jamming attacks without relying on the known jamming strategies. The deep RL based rate adaption scheme in [27] reduces the packet loss rate and transmission latency by using the recent dueling neural network and the ambient backscattering against the reactive and smart jammers. The underwater relay node in [28] that helps the sensor forward the acoustic signals applies Q-learning and deep Q-network to determine its location and relay power, resist a smart jammer and achieve the BER and energy consumption lower bounds. The cellular system in [29] uses a UAV as the relay to resist the Q-learning based jamming attacks and combines deep Q-network with transfer learning to optimize the UAV transmit power with reduced BER and communication overhead, and less energy consumption.

3 System model

As illustrated in Figure 1, a UAV with multiple sensors such as cameras collects the sensing data such as the traffic information and sends the data to the BS for traffic management with power $P_{\rm U}$ and channel gain $h_{\rm U,B}^{(k)}$ at time slot k. A robot located at $(x_1^{(k-1)}, x_2^{(k-1)})$ m receives the UAV message with channel gain $h_{\rm U,R}^{(k)}$ and RSSI $r^{(k)}$ and helps relay the message if the UAV-BS link is interrupted by a jammer.

The robot moves $a_1^{(k)} \in [-X, X]$ m along the x-axis, $a_2^{(k)} \in [-Y, Y]$ m along the y-axis to location $(x_1^{(k-1)} + a_1^{(k)}, x_2^{(k-1)} + a_2^{(k)})$ m, and then relays the message to the BS with power $a_3^{(k)}$ and channel gain $h_{\rm R,B}^{(k)}$, where X and Y are the maximum moving distance along the two axes. For simplicity, the robot is assumed to have L_1 and L_2 feasible moving distances along the x-axis and y-axis, respectively. The relay power with maximum constraint $P_{\rm R}$ is quantized into L_3 levels, i.e., $a_3^{(k)} \in \{iP_{\rm R}/L_3 | 0 \leqslant i \leqslant L_3\}$.

The BS applies the combining algorithm as designed in [30] that provides an accurate BER estimation without requiring to know the channel state information to combine the messages from the robot and the UAV, and uses the cumulative distribution function of the received signals to measure the BER $b^{(k)}$

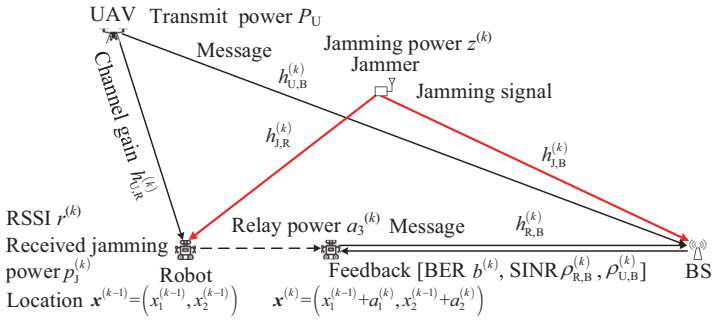

Figure 1 (Color online) Illustration of a robot aided UAV-ground communication system, in which the robot chooses its moving distance $a_1^{(k)}$ and $a_2^{(k)}$, and relay power $a_3^{(k)}$ to help forward the UAV message to the BS against a smart jammer.

Table 1 Summary of symbols and notations

Symbol	Description
P_{U}	UAV transmit power
X/Y	Maximum moving distance of the robot along the x-axis/y-axis
$L_{1/2}$	Feasible moving distances along the x-axis/y-axis
$a_1^{(k)} \in [-X, X]$ /$a_2^{(k)} \in [-Y, Y]$	Moving distance along the x/y-axis at time slot k
$P_{\mathrm{R/J}}$	Maximum relay power/jamming power
L_3	Feasible relay power levels
$a_3^{(k)} \in \{\frac{iP_{\mathrm{R}}}{L_3} \mid 0 \leqslant i \leqslant L_3\}$	Relay power
$z^{(k)} \in [0, P_{\mathrm{J}}]$	Jamming power
$h^{(k)}_{\mathrm{U/R,B}}$	Channel gain of the UAV/robot-BS link
$h^{(k)}_{\mathrm{U,R}}$	Channel gain of the UAV-robot link
$h^{(k)}_{\mathrm{J,R/B}}$	Channel gain of the jammer-robot/BS link
$r^{(k)}$	RSSI of the signal received by the robot
$p^{(k)}_{\mathrm{J}}$	Jamming power received by the robot
$b^{(k)}$	BER of the UAV messages
$\rho^{(k)}_{\mathrm{U/R,B}}$	SINR of the signal received by the BS from the UAV/robot
$\rho^{(k)}_{\mathrm{U,R}}$	SINR of the signal received by the robot from the UAV
$\varsigma^{(k)}$	Robot energy consumption

based on the given modulation and channel coding scheme. The SINR of the UAV signal $\rho^{(k)}_{\mathrm{U,B}}$ and that of the robot signal $\rho^{(k)}_{\mathrm{R,B}}$ are estimated based on $b^{(k)}$, which are sent to the robot. For simplicity, both the robot and the BS are assumed to receive the noise signals with the resulting power given by σ^2.

Jammers aim to degrade the transmission quality of the UAV messages, exhaust the UAV energy, and even launch the denial of service attacks by sending jamming signals. In particular, a smart jammer uses a universal software radio peripheral to continuously eavesdrop the channel transmission status, and then flexibly changes its jamming power to attack the UAV transmission accordingly [4]. More specifically, the smart jammer that has a maximum power given by P_{J} applies RL to choose the jamming power $z^{(k)} \in [0, P_{\mathrm{J}}]$ and sends the jamming signal to the robot and the BS.

The channel gain of the jammer-BS link is denoted as $h^{(k)}_{\mathrm{J,B}}$. The robot receives the jamming signal with received power $p^{(k)}_{\mathrm{J}}$ and channel gain $h^{(k)}_{\mathrm{J,R}}$, and then estimates the SINR of the UAV signal denoted as $\rho^{(k)}_{\mathrm{U,R}}$ based on the RSSI $r^{(k)}$ and the received jamming power $p^{(k)}_{\mathrm{J}}$. Some important symbols and notations are summarized in Table 1, and the time index k is omitted if no confusions occur.

4 RL-based robot relay scheme

We propose an RL-based robot relay scheme (RLRR) to determine the mobility and relay policy $\boldsymbol{a}^{(k)} = [a_i^{(k)}]_{1\leqslant i\leqslant 3}$. This scheme combines reinforcement learning with a tile coding technique to calculate the Q-values for each mobility and relay policy under the current state denoted as $\boldsymbol{o}^{(k)}$, instead of using the iterative Bellman equation to update the Q-values for higher storage efficiency. More specifically, this scheme applies the tile coding technique to map all the feasible states into M tilings, and each tiling consists of C tiles, with $C > M$.

The robot measures the received jamming power $p_{\mathrm{J}}^{(k)}$ and the RSSI $r^{(k)}$, and observes its location $(x_1^{(k-1)}, x_2^{(k-1)})$ at time slot k, which are used to form the state. The state $\boldsymbol{o}^{(k)}$ also relies on the previous energy consumption denoted as $\varsigma^{(k-1)}$ and the previous BER $b^{(k-1)}$ obtained from the BS, which is given by

$$\boldsymbol{o}^{(k)} = \left[p_{\mathrm{J}}^{(k)}, r^{(k)}, \left(x_1^{(k-1)}, x_2^{(k-1)}\right), \varsigma^{(k-1)}, b^{(k-1)}\right] \in \Omega, \tag{1}$$

where the continuous state space Ω consists of all the available jamming power levels and the feasible RSSIs, moving distances, energy consumption levels and BER levels.

Similar to [15], this scheme uses the hash function to map the state $\boldsymbol{o}^{(k)}$ into M active tiles, denoted as $\{f(\boldsymbol{o}^{(k)}, \boldsymbol{a}', i)\}_{1 \leqslant i \leqslant M}$, for each mobility and relay policy $\boldsymbol{a}' \in \boldsymbol{A}$, where $f(\boldsymbol{o}^{(k)}, \boldsymbol{a}', i) \in \{1, 2, \ldots, CM\}$ is the i-th active tile of the state-action pair $(\boldsymbol{o}^{(k)}, \boldsymbol{a}')$, and the action set $\boldsymbol{A}$ consists of the $L_1 L_2 (L_3 + 1)$ feasible mobility and relay policies.

The weight value denoted as $\boldsymbol{\omega}(\cdot)$ is the expected utility in the relay process, with each state-action pair having M weight values. More specifically, the weight value of the i-th active tile under $(\boldsymbol{o}^{(k)}, \boldsymbol{a}')$ is given by $\boldsymbol{\omega}(f(\boldsymbol{o}^{(k)}, \boldsymbol{a}', i))$. Instead of directly updating the Q-values denoted as $Q(\boldsymbol{o}^{(k)}, \cdot)$ via the iterative Bellman equation, this scheme uses the M weight values to compute the Q-values as follows:

$$Q\left(\boldsymbol{o}^{(k)}, \boldsymbol{a}'\right) = \sum_{i=1}^{M} \boldsymbol{w}\left(f\left(\boldsymbol{o}^{(k)}, \boldsymbol{a}', i\right)\right), \quad \forall \boldsymbol{a}' \in \boldsymbol{A}. \tag{2}$$

Based on the Q-values $Q(\boldsymbol{o}^{(k)}, \cdot)$ and ϵ-greedy algorithm, this scheme chooses the mobility and relay policy $\boldsymbol{a}^{(k)}$. As a result, the robot moves to location $(x_1^{(k-1)} + a_1^{(k)}, x_2^{(k-1)} + a_2^{(k)})$, and relays the UAV message to the BS with power $a_3^{(k)}$. Upon receiving the feedback from the BS, the robot obtains the BER $b^{(k)}$, and the SINR $\rho_{\mathrm{U,B}}^{(k)}$ and $\rho_{\mathrm{R,B}}^{(k)}$. The robot measures the energy consumption $\varsigma^{(k)}$ and estimates the SINR of the received signal from the UAV $\rho_{\mathrm{U,R}}^{(k)}$ based on the RSSI $r^{(k)}$ and the received jamming power $p_{\mathrm{J}}^{(k)}$ to compute the utility by

$$u_{\mathrm{R}}^{(k)} = \max\left\{\min\left\{\rho_{\mathrm{U,R}}^{(k)}, \rho_{\mathrm{R,B}}^{(k)}\right\}, \rho_{\mathrm{U,B}}^{(k)}\right\} - c_1 \varsigma^{(k)}, \tag{3}$$

where $c_1 > 0$ is the coefficient that represents the importance of the energy consumption in the utility evaluation. Let $\alpha \in (0, 1]$ denote the learning rate and $\beta \in [0, 1]$ be the discount factor. As shown in Algorithm 1, this scheme uses the iterative Bellman equation to update the M weight values $\boldsymbol{\omega}(f(\boldsymbol{o}^{(k-1)}, \boldsymbol{a}^{(k-1)}, \cdot))$ similar to [15].

Algorithm 1 RL-based robot relay scheme

Initialize M, C, β, α, $\boldsymbol{A}$, $\boldsymbol{o}^{(0)}$, $\boldsymbol{a}^{(0)}$, $\boldsymbol{\omega} = \boldsymbol{0}$, $\varsigma^{(0)}$, $\boldsymbol{x}^{(0)}$ and $b^{(0)}$;
for $k = 1, 2, \ldots$ **do**
  Measure $p_{\mathrm{J}}^{(k)}$ and $r^{(k)}$;
  Observe $(x_1^{(k-1)}, x_2^{(k-1)})$;
  Form $\boldsymbol{o}^{(k)}$ via (1);
  for $\boldsymbol{a}' \in \boldsymbol{A}$ **do**
    Map $\boldsymbol{o}^{(k)}$ into M active tiles, $\{f(\boldsymbol{o}^{(k)}, \boldsymbol{a}', i)\}_{1 \leqslant i \leqslant M}$;
    Calculate $Q(\boldsymbol{o}^{(k)}, \boldsymbol{a}')$ via (2);
  end for
  Choose $\boldsymbol{a}^{(k)}$ with ϵ-greedy based on $Q(\boldsymbol{o}^{(k)}, \cdot)$
  Move to location $(x_1^{(k-1)} + a_1^{(k)}, x_2^{(k-1)} + a_2^{(k)})$;
  Relay the UAV message with power $a_3^{(k)}$;
  Receive the feedback from the BS;
  Obtain $b^{(k)}$, $\rho_{\mathrm{R,B}}^{(k)}$ and $\rho_{\mathrm{U,B}}^{(k)}$;
  Measure $\varsigma^{(k)}$;
  Estimate $\rho_{\mathrm{U,R}}^{(k)}$;
  Calculate $u_{\mathrm{R}}^{(k)}$ via (3);
  for $i = 1, 2, \ldots, M$ **do**
    $\boldsymbol{\omega}\left(f\left(\boldsymbol{o}^{(k-1)}, \boldsymbol{a}^{(k-1)}, i\right)\right) \leftarrow \boldsymbol{\omega}\left(f\left(\boldsymbol{o}^{(k-1)}, \boldsymbol{a}^{(k-1)}, i\right)\right) + \alpha\left(u_{\mathrm{R}}^{(k-1)} + \beta \max_{\boldsymbol{a}' \in \boldsymbol{A}} Q\left(\boldsymbol{o}^{(k)}, \boldsymbol{a}'\right) - Q\left(\boldsymbol{o}^{(k-1)}, \boldsymbol{a}^{(k-1)}\right)\right)$;
  end for
end for

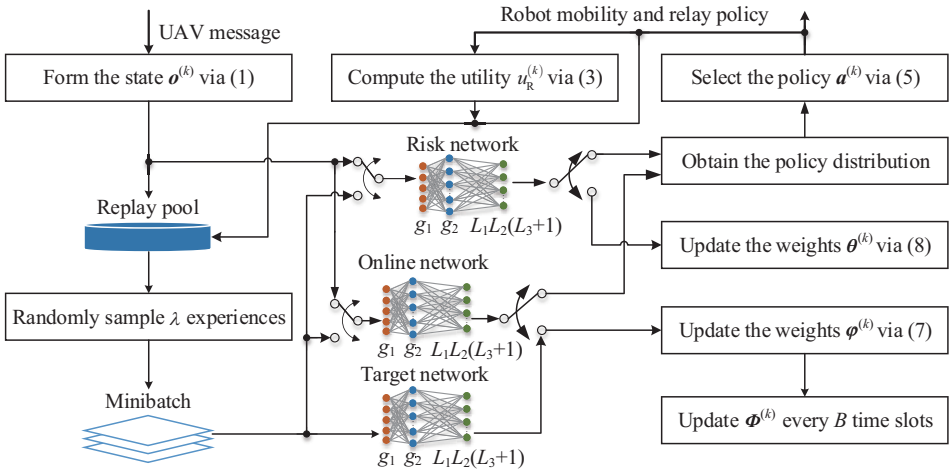

Figure 2 (Color online) Illustration of the deep RL-based robot relay scheme.

5 Deep RL-based robot relay scheme

We propose a deep RL-based relay scheme (DRLRR) with safe exploration for the robot that has sufficient computing resources to resist jamming. This scheme uses three deep neural networks with the same network architecture to avoid exploring the dangerous mobility and relay policies and reduce the overestimation of the Q-values, and uses fully connected layers instead of convolutional layers for lower computational complexity. More specifically, an online network outputs the Q-values, a target network evaluates the Q-values of the chosen mobility and relay policy, and a risk network outputs the long-term risk values.

Similar to Algorithm 1, the five-dimensional state $\boldsymbol{o}^{(k)}$ is formulated via (1) and input into the online network with weights $\boldsymbol{\varphi}^{(k-1)}$ and the risk network with weights $\boldsymbol{\theta}^{(k-1)}$. This scheme uses the BER $b(\boldsymbol{o}^{(k)}, \boldsymbol{a}^{(k)})$ to evaluate the risk of choosing the policy $\boldsymbol{a}^{(k)} \in \boldsymbol{A}$ under state $\boldsymbol{o}^{(k)}$. If the BER $b(\boldsymbol{o}^{(k)}, \boldsymbol{a}^{(k)})$ is larger than the quality of service threshold defined as ξ, the risk value of state-action pair $(\boldsymbol{o}^{(k)}, \boldsymbol{a}^{(k)})$ defined as $\boldsymbol{I}(\boldsymbol{o}^{(k)}, \boldsymbol{a}^{(k)})$ is set as 1, and 0 otherwise. Let $\gamma \in [0, 1]$ be the discount coefficient that parametrizes the uncertainty regarding the future risk values. The long-term risk values can be defined as $R(\boldsymbol{o}^{(k)}, \boldsymbol{a}^{(k)}; \boldsymbol{\theta}^{(k-1)})$ which depend on the future N risk values and the weights of the risk network $\boldsymbol{\theta}^{(k-1)}$:

$$R\left(\boldsymbol{o}^{(k)}, \boldsymbol{a}^{(k)}; \boldsymbol{\theta}^{(k-1)}\right) = \sum_{i=0}^{N} \gamma^i \boldsymbol{I}\left(b\left(\boldsymbol{o}^{(k)}, \boldsymbol{a}^{(k)}\right) > \xi\right). \tag{4}$$

As shown in Figure 2, the risk network has three fully connected layers, including an input layer with g_1 neurons, a hidden layer with g_2 neurons, and an output layer with $L_1L_2(L_3+1)$ neurons. The hidden layer is activated by rectified linear units (ReLU) and the output layer activated by the softmax function obtains $L_1L_2(L_3+1)$ risk values of the feasible mobility and relay policies, i.e., $R(\boldsymbol{o}^{(k)}, \cdot; \boldsymbol{\theta}^{(k-1)})$. The online network has the same architecture as the risk network, uses ReLU as the activation function in the hidden layer, and outputs $L_1L_2(L_3+1)$ Q-values $Q(\boldsymbol{o}^{(k)}, \cdot; \boldsymbol{\varphi}^{(k-1)})$.

This scheme uses Q-values and long-term risk values as the basis to choose the mobility and relay policy $\boldsymbol{a}^{(k)}$, with the policy distribution given by

$$\Pr\left(\boldsymbol{a}^{(k)} = \tilde{\boldsymbol{a}}\right) = \frac{\exp\left(\frac{Q(\boldsymbol{o}^{(k)}, \tilde{\boldsymbol{a}}; \boldsymbol{\varphi}^{(k-1)})}{1+R(\boldsymbol{o}^{(k)}, \tilde{\boldsymbol{a}}; \boldsymbol{\theta}^{(k-1)})}\right)}{\sum_{\hat{\boldsymbol{a}} \in \boldsymbol{A}} \exp\left(\frac{Q(\boldsymbol{o}^{(k)}, \hat{\boldsymbol{a}}; \boldsymbol{\varphi}^{(k-1)})}{1+R(\boldsymbol{o}^{(k)}, \hat{\boldsymbol{a}}; \boldsymbol{\theta}^{(k-1)})}\right)}. \tag{5}$$

According to the chosen $\boldsymbol{a}^{(k)}$, the robot moves to the new location $(x_1^{(k-1)} + a_1^{(k)}, x_2^{(k-1)} + a_2^{(k)})$ and relays the message with power $a_3^{(k)}$. This scheme measures the robot energy consumption $\varsigma^{(k)}$ and estimates the SINR $\rho_{\mathrm{U,R}}^{(k)}$ to calculate the utility $u_{\mathrm{R}}^{(k)}$ via (3) similar to Algorithm 1.

This scheme saves the previous state-action pair $(\boldsymbol{o}^{(k-1)}, \boldsymbol{a}^{(k-1)})$, the previous utility $u_{\mathrm{R}}^{(k-1)}$, the current

state $\boldsymbol{o}^{(k)}$, and the previous risk value $\boldsymbol{I}(b(\boldsymbol{o}^{(k-1)}, \boldsymbol{a}^{(k-1)}) > \xi)$ as the robot relay experience given by

$$\vartheta^{(k)} = \left\{ \boldsymbol{o}^{(k-1)}, \boldsymbol{a}^{(k-1)}, u_{\mathrm{R}}^{(k-1)}, \boldsymbol{o}^{(k)}, \boldsymbol{I}\left(b\left(\boldsymbol{o}^{(k-1)}, \boldsymbol{a}^{(k-1)}\right) > \xi\right) \right\}, \tag{6}$$

which is input to the replay pool defined as $\mathcal{M}$. By randomly sampling λ relay experiences $[\vartheta^{(m(j))}]_{1 \leqslant j \leqslant \lambda}$ from $\mathcal{M}$, this scheme applies the stochastic gradient decent algorithm in [16] to update the weights $\boldsymbol{\varphi}^{(k)}$ and $\boldsymbol{\theta}^{(k)}$, with $m(j) \sim U(1, k)$. With the same network architecture in the online network, a target network with weights $\boldsymbol{\phi}^{(k-1)}$ uses the sampled λ relay experiences to estimate the target Q-values defined as $Q(\boldsymbol{o}^{(k)}, \cdot; \boldsymbol{\phi}^{(k-1)})$ and update the online network weights $\boldsymbol{\varphi}^{(k)}$ [31]. More specifically, the online network weights $\boldsymbol{\varphi}^{(k)}$ are updated to minimize the mean square error between the Q-values and the target Q-values as follows:

$$\begin{aligned} \boldsymbol{\varphi}^{(k)} = \arg\min_{\boldsymbol{\varphi}'} \frac{1}{\lambda} \sum_{j=1}^{\lambda} \Bigg(& u_{\mathrm{R}}^{(m(j)-1)} + \beta Q\left(\boldsymbol{o}^{(m(j))}, \arg\max_{\boldsymbol{a}' \in \boldsymbol{A}} Q\left(\boldsymbol{o}^{(m(j))}, \boldsymbol{a}'; \boldsymbol{\varphi}^{(k-1)}\right); \boldsymbol{\phi}^{(k-1)}\right) \\ & - Q\left(\boldsymbol{o}^{(m(j)-1)}, \boldsymbol{a}^{(m(j)-1)}; \boldsymbol{\varphi}'\right) \Bigg)^2. \end{aligned} \tag{7}$$

The risk network weights $\boldsymbol{\theta}^{(k)}$ are updated with the stochastic gradient decent algorithm as follows:

$$\begin{aligned} \boldsymbol{\theta}^{(k)} = \arg\min_{\boldsymbol{\theta}'} \frac{1}{\lambda} \sum_{j=1}^{\lambda} \Bigg(& \boldsymbol{I}\left(b\left(\boldsymbol{o}^{(m(j)-1)}, \boldsymbol{a}^{(m(j)-1)}\right) > \xi\right) + \beta \min_{\boldsymbol{a}' \in \boldsymbol{A}} R\left(\boldsymbol{o}^{(m(j))}, \boldsymbol{a}'; \boldsymbol{\theta}^{(k-1)}\right) \\ & - R\left(\boldsymbol{o}^{(m(j)-1)}, \boldsymbol{a}^{(m(j)-1)}; \boldsymbol{\theta}'\right) \Bigg)^2. \end{aligned} \tag{8}$$

The target network weights $\boldsymbol{\phi}^{(k)}$ are updated with the online network weights $\boldsymbol{\varphi}^{(k)}$ every B time slots to avoid the instability exploration, as shown in Algorithm 2.

Algorithm 2 Deep RL-based robot relay scheme

Initialize γ, β, $\boldsymbol{A}$, $\boldsymbol{o}^{(0)}$, $\boldsymbol{a}^{(0)}$, $\varsigma^{(0)}$, $\boldsymbol{x}^{(0)}$, $b^{(0)}$, N, λ, B, $\mathcal{M} = \emptyset$, $\boldsymbol{\varphi}^{(0)}$, $\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\phi}^{(0)}$;
for $k = 1, 2, \ldots$ **do**
  Measure $p_{\mathrm{J}}^{(k)}$ and $r^{(k)}$;
  Observe $(x_1^{(k-1)}, x_2^{(k-1)})$;
  Form $\boldsymbol{o}^{(k)}$ via (1);
  Input $\boldsymbol{o}^{(k)}$ to the online network and the risk network;
  Risk network outputs, $R(\boldsymbol{o}^{(k)}, \cdot; \boldsymbol{\theta}^{(k-1)})$;
  Online network outputs, $Q(\boldsymbol{o}^{(k)}, \cdot; \boldsymbol{\varphi}^{(k-1)})$;
  Choose $\boldsymbol{a}^{(k)}$ via (5);
  Move to location $(x_1^{(k-1)} + a_1^{(k)}, x_2^{(k-1)} + a_2^{(k)})$;
  Relay the UAV message with power $a_3^{(k)}$;
  Receive the feedback from the BS;
  Obtain $b^{(k)}$, $\rho_{\mathrm{R,B}}^{(k)}$ and $\rho_{\mathrm{U,B}}^{(k)}$;
  Measure $\varsigma^{(k)}$;
  Estimate $\rho_{\mathrm{U,R}}^{(k)}$;
  Calculate $u_{\mathrm{R}}^{(k)}$ via (3);
  Save $\vartheta^{(k)}$ via (6);
  $\mathcal{M} = \mathcal{M} \cup \vartheta^{(k)}$;
  if $k > \lambda$ **then**
    Randomly choose λ relay experiences from $\mathcal{M}$, $[\vartheta^{(m(j))}]_{1 \leqslant j \leqslant \lambda}$;
    Update $\boldsymbol{\varphi}^{(k)}$ via (7);
    Update $\boldsymbol{\theta}^{(k)}$ via (8);
    if $k \mod B = 0$ **then**
      $\boldsymbol{\phi}^{(k)} \leftarrow \boldsymbol{\varphi}^{(k)}$;
    end if
  end if
end for

6 Performance analysis

We provide the computational complexity, the lower bounds including the BER and the energy consumption, and the utility upper bound of the proposed robot relay schemes. More specifically, the robot

relay process is formulated as a robot relay anti-jamming game, where the robot chooses the moving policy $a_1 \in [-X, X]$ and $a_2 \in [-Y, Y]$ and its relay power $a_3 \in [0, P_{\mathrm{R}}]$, and the jammer selects its power $z \in [0, P_{\mathrm{J}}]$. Both the robot and the jammer aim to maximize their utility with reduced energy consumption.

For simplicity, the channel gains of the UAV-BS/robot link, jammer-BS/robot link, and robot-BS link are assumed to be a constant. The robot energy consumption consists of the transmission and moving energy consumption, which is modeled as

$$\varsigma = c_2 a_3 + c_3 \sqrt{a_1^2 + a_2^2}, \tag{9}$$

where c_2 is the unit transmission energy consumption, and c_3 is the unit moving energy consumption.

By (3) and (9), we have the robot utility given by

$$u_{\mathrm{R}}(\boldsymbol{a}, z) = \max\left\{\min\left\{\frac{P_{\mathrm{U}} h_{\mathrm{U,R}}}{z h_{\mathrm{J,R}} + \sigma^2}, \frac{a_3 h_{\mathrm{R,B}}}{z h_{\mathrm{J,B}} + \sigma^2}\right\}, \frac{P_{\mathrm{U}} h_{\mathrm{U,B}}}{z h_{\mathrm{J,B}} + \sigma^2}\right\} - c_1 c_2 a_3 - c_1 c_3 \sqrt{a_1^2 + a_2^2}. \tag{10}$$

The UAV message is assumed to use the binary phase-shift keying in the transmission, and the resulting BER is given by

$$b = \frac{1}{2}\mathrm{erfc}\left(\sqrt{\max\left\{\min\left\{\frac{P_{\mathrm{U}} h_{\mathrm{U,R}}}{z h_{\mathrm{J,R}} + \sigma^2}, \frac{a_3 h_{\mathrm{R,B}}}{z h_{\mathrm{J,B}} + \sigma^2}\right\}, \frac{P_{\mathrm{U}} h_{\mathrm{U,B}}}{z h_{\mathrm{J,B}} + \sigma^2}\right\}}\right). \tag{11}$$

By choosing the jamming power z ranging from zero to P_{J}, the jammer aims to degrade the robot utility and improve its own utility defined as u_{J}, which is modeled based on the robot utility u_{R} and the jamming energy consumption as follows:

$$u_{\mathrm{J}}(\boldsymbol{a}, z) = -u_{\mathrm{R}} - c_4 z, \tag{12}$$

where c_4 is the unit jamming energy consumption.

Theorem 1. The bounds of the RL based robot relay scheme are given by

$$\varsigma = c_2 P_{\mathrm{R}}, \tag{13}$$

$$b = \frac{1}{2}\mathrm{erfc}\left(\sqrt{\frac{P_{\mathrm{R}} h_{\mathrm{R,B}}}{P_{\mathrm{J}} h_{\mathrm{J,B}} + \sigma^2}}\right), \tag{14}$$

$$u_{\mathrm{R}} = \frac{P_{\mathrm{R}} h_{\mathrm{R,B}}}{P_{\mathrm{J}} h_{\mathrm{J,B}} + \sigma^2} - c_1 c_2 P_{\mathrm{R}}, \tag{15}$$

if

$$\max\left\{c_1 c_2\left(P_{\mathrm{J}} h_{\mathrm{J,B}} + \sigma^2\right), \frac{P_{\mathrm{U}} h_{\mathrm{U,B}}(P_{\mathrm{J}} h_{\mathrm{J,B}} + \sigma^2)}{P_{\mathrm{R}}}, \frac{c_4 (P_{\mathrm{J}} h_{\mathrm{J,B}} + \sigma^2)^2}{P_{\mathrm{R}} h_{\mathrm{J,B}}}\right\} \leqslant h_{\mathrm{R,B}} < \frac{P_{\mathrm{U}} h_{\mathrm{U,R}}}{P_{\mathrm{R}}(P_{\mathrm{J}} h_{\mathrm{J,R}} + \sigma^2)}. \tag{16}$$

Proof. By (10) and (12), if Eq. (16) holds, $\forall z \in [0, P_{\mathrm{J}}]$, we have

$$\frac{\partial^2 u_{\mathrm{J}}([0, 0, P_{\mathrm{R}}], z)}{\partial z^2} = -\frac{2 P_{\mathrm{R}} h_{\mathrm{R,B}} h_{\mathrm{J,B}}^2}{\left(z h_{\mathrm{J,B}} + \sigma^2\right)^3} < 0, \tag{17}$$

$$\frac{\partial u_{\mathrm{J}}([0, 0, P_{\mathrm{R}}], z)}{\partial z} = \frac{P_{\mathrm{R}} h_{\mathrm{R,B}} h_{\mathrm{J,B}}}{\left(z h_{\mathrm{J,B}} + \sigma^2\right)^2} - c_4 \geqslant 0. \tag{18}$$

Thus, we have

$$u_{\mathrm{J}}([0, 0, P_{\mathrm{R}}], P_{\mathrm{J}}) \geqslant u_{\mathrm{J}}([0, 0, P_{\mathrm{R}}], z). \tag{19}$$

Let

$$n(a_1, a_2) = -c_1 c_3 \sqrt{a_1^2 + a_2^2}, \tag{20}$$

and $\forall a_1 \in [-X, X]$ and $a_2 \in [-Y, Y]$, we have

$$n(0,0) \geqslant n(a_1, a_2). \tag{21}$$

Thus, $n(a_1, a_2)$ is maximized at $(a_1^*, a_2^*) = (0,0)$. Let

$$l(a_3) = \max\left\{\min\left\{\frac{P_{\mathrm{U}} h_{\mathrm{U,R}}}{P_{\mathrm{J}} h_{\mathrm{J,R}} + \sigma^2}, \frac{a_3 h_{\mathrm{R,B}}}{P_{\mathrm{J}} h_{\mathrm{J,B}} + \sigma^2}\right\}, \frac{P_{\mathrm{U}} h_{\mathrm{U,B}}}{P_{\mathrm{J}} h_{\mathrm{J,B}} + \sigma^2}\right\} - c_1 c_2 a_3, \tag{22}$$

and $\forall a_3 \in [0, P_{\mathrm{R}}]$, if Eq. (16) holds, we have

$$\frac{\mathrm{d}l(a_3)}{\mathrm{d}a_3} = \frac{h_{\mathrm{R,B}}}{P_{\mathrm{J}} h_{\mathrm{J,B}} + \sigma^2} - c_1 c_2 \geqslant 0. \tag{23}$$

Thus, $l(a_3)$ is maximized at $a_3^* = P_{\mathrm{R}}$. By (10), (20) and (22), we have

$$u_{\mathrm{R}}(\boldsymbol{a}, P_{\mathrm{J}}) = n(a_1, a_2) + l(a_3). \tag{24}$$

As both a_1 and a_2 are independent with a_3, we have

$$u_{\mathrm{R}}([0, 0, P_{\mathrm{R}}], P_{\mathrm{J}}) \geqslant u_{\mathrm{R}}(\boldsymbol{a}, P_{\mathrm{J}}). \tag{25}$$

Thus, by (19) and (25), we have that $([0, 0, P_{\mathrm{R}}], P_{\mathrm{J}})$ is a Nash equilibrium of the robot relay game. Hence, according to (9)–(11), we have the bounds given by (13)–(15).

Remark 1. The robot stays in a fixed location and relays the UAV message to the BS with its maximum power P_{R} against the jammer who sends jamming signals with maximum power P_{J}, if the channel gain of the robot-BS link satisfies the bounds that depend on the transmit power of the UAV, the robot and the jammer, as shown in (16). The robot energy consumption given by (13) only relies on the relay power. Both the robot maximum relay power and the maximum jamming power affect the BER of the UAV messages given by (14).

The computational complexity of DRLRR relies on the complexity of the three networks in the forward and backward propagations. Both the online network and the risk network update their weights with the forward and backward propagations, and the target network updates the weights with the backward propagation.

According to [32], the number of the multiplications in the forward propagation defined as ε_1 relies on the input size 5, the output size $L_1 L_2 (L_3 + 1)$ and the sample relay experience size λ, and is given by

$$\varepsilon_1 = 6\lambda g_1 + \lambda (g_1 + 1) g_2 + \lambda L_1 L_2 (g_2 + 1)(L_3 + 1). \tag{26}$$

Similarly, the number of the multiplications in the backward propagation defined as ε_2 is given by

$$\varepsilon_2 = 12\lambda g_1 + 2\lambda (g_1 + 1) g_2 + 3\lambda L_1 L_2 (g_2 + 1)(L_3 + 1). \tag{27}$$

Theorem 2. The computational complexity of QTC in [11] at time slot k is $O(L_1 L_2 L_3)$, that of RLRR in Algorithm 1 is $O(M L_1 L_2 L_3)$, and that of DRLRR in Algorithm 2 is given by

$$\psi = O(g_2 \lambda L_1 L_2 L_3). \tag{28}$$

Proof. According to [33], QTC in [11] has the computational complexity given by $O(L_1 L_2 L_3)$. Compared with QTC, our proposed RLRR scheme maps the state into M active tiles to compute the Q-values in each time slot. Thus, the computational complexity of RLRR in Algorithm 1 is $O(M L_1 L_2 L_3)$.

According to [32], the computational complexity of the proposed DRLRR in Algorithm 2 at time slot k is given by

$$\begin{aligned}
\psi &= O(2\varepsilon_1 + 3\varepsilon_2) \\
&= O(\lambda(48 g_1 + 8(g_1 + 1) g_2 + 11(g_2 + 1)(L_1 + 1) L_2 L_3)) && (29) \\
&= O(\lambda(g_1 + (g_1 + 1) g_2 + (g_2 + 1)(L_1 + 1) L_2 L_3)) && (30) \\
&= O(g_1 \lambda + g_1 g_2 \lambda + g_2 \lambda L_1 L_2 L_3) && (31) \\
&= O(g_1 g_2 \lambda + g_2 \lambda L_1 L_2 L_3) && (32) \\
&= O(g_2 \lambda L_1 L_2 L_3). && (33)
\end{aligned}$$

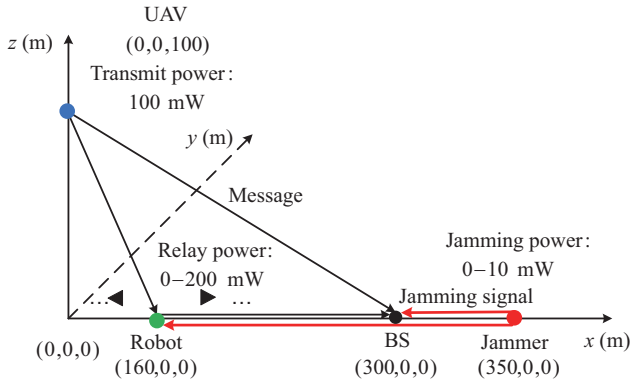

Figure 3 (Color online) Simulation settings of a robot aided UAV-ground communication system against a jammer located at (350, 0, 0) m that chooses the jamming power from {0, 2.5, 5, 7.5, 10} mW, in which a robot forwards the UAV messages with relay power ranging from zero to 200 mW.

Remark 2. The computational complexity of RLRR increases with the number of the exploration samples, the number of the mapped active tiles, and the number of feasible mobility and relay policies of the robot. The computational complexity of DRLRR also depends on the sample relay experience size and the number of neurons in the hidden layer. DRLRR in Algorithm 2 that has higher computational complexity than RLRR enables the robot with enough computing resources to optimize the mobility and relay policy with reduced sample complexity.

7 Simulation results

We perform simulations based on a typical UAV-ground communication system at the center frequency 2.452 GHz to evaluate the performance of the robot relay schemes, with the initial topology as shown in Figure 3. A three-dimensional coordinate system is considered, and the robot is assumed to only move along the x-axis in the simulations. A UAV located at $(0, 0, 100)$ m uses the camera sensor to collect the traffic information and aims to send the collected data to the BS located at $(300, 0, 0)$ m with power 100 mW and frequency 2.452 GHz [34], whose channel gain changes between 3.5×10^{-13} and 4.0×10^{-13}.

The robot initially located at (160, 0, 0) m chooses the moving distance from zero to 5 m and then moves to a new location with 2 mJ unit energy consumption. The channel gain from the UAV to the robot changes between 1.6×10^{-12} and 2.5×10^{-11}. According to the practical robots following SRRC such as RoboMaster S1, the robot relays the UAV messages with power chosen from zero to 200 mW and has 8 levels [35]. The channel gain of the robot-BS link is assumed to follow a log-distance path loss model with exponent 3 according to [18], and the transmission cost coefficient is 0.005. According to [36], the quality of service threshold is chosen as 10^{-3}.

A smart jammer at a fixed location $(350, 0, 0)$ m uses a universal software radio peripheral to eavesdrop on the transmission channel status of the UAV and applies Q-learning to choose its power from $\{0, 2.5, 5, 7.5, 10\}$ mW with jamming cost coefficient 0.01. The jammer sends jamming signals to the robot and the BS with both the channel gains following the log-distance path loss channel model in [18], and the corresponding exponents are given by 3 and 3.9, respectively. The noise power received by both the robot and the jammer is 10^{-14} mW according to [34].

The learning parameters are chosen according to the simulated training results not shown here with reduced BER, outage probability and energy consumption. More specifically, the learning rate and the discount factor are chosen as 0.8 and 0.6, respectively. The feasible states are mapped into 16 tilings, with each tiling consisting of 1280 tiles. The robot randomly chooses 64 relay experiences in each time slot to update the weights of both the online and risk networks, and updates the target network weights every 50 time slots. The three fully connected layers of each network have 5, 256 and 27 neurons, respectively. In the simulations, each time slot lasts 6 s.

As shown in Figure 4, the BER of the UAV message, the outage probability and the robot energy consumption decrease with time, and the robot utility increases with time. For example, RLRR decreases the BER from 5.5‰ to 1.2‰, reduces the outage probability from 79.3% to 32.0%, and saves the robot

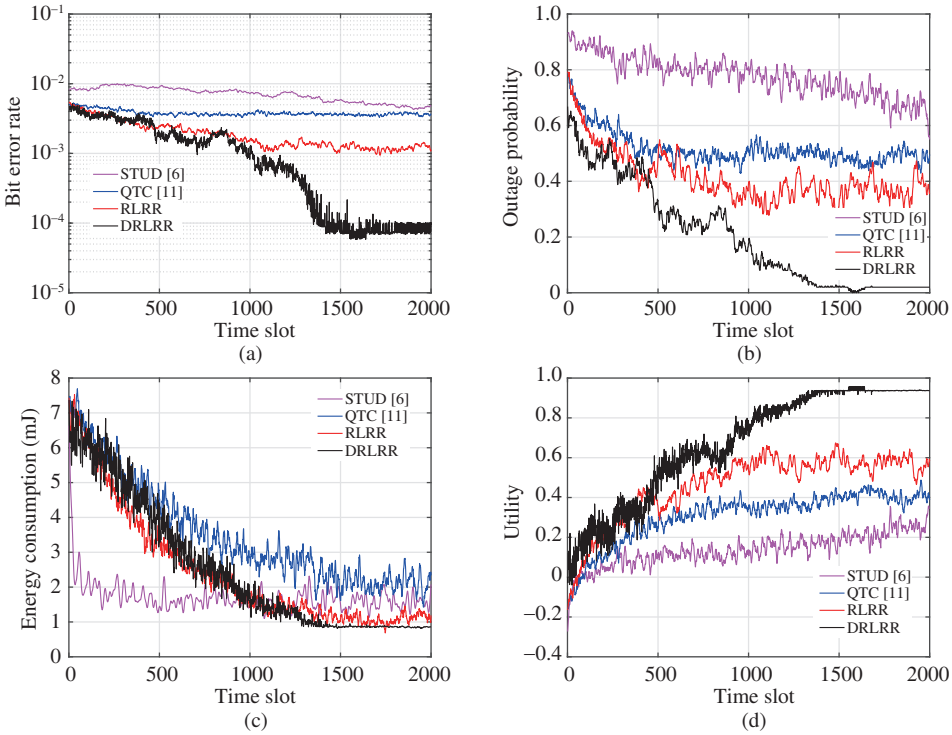

Figure 4 (Color online) Performance of the relay schemes in a robot aided UAV-ground communication system against a smart jammer whose maximum jamming power is 10 mW as shown in Figure 3. (a) BER; (b) outage probability; (c) energy consumption; (d) utility.

energy from 7.4 to 1.1 mJ after 1500 time slots, which satisfies the quality of service requirement for a typical UAV-ground communication system in [36].

Our proposed relay schemes outperform the benchmark QTC in [11] by reducing the BER and the outage probability, saving the robot energy and increasing the utility. For example, the proposed RLRR reduces the BER by 67.5%, the outage probability by 36.6%, and the robot energy by 31.4% and improves the utility by 50.7% at the 1500-th time slot compared with QTC. In addition, RLRR exceeds STUD in [6] with 79.5% lower BER, 56.4% lower outage probability, and 26.1% less energy consumption after 1500 time slots. That is because RLRR uses a robot as a relay to provide a better transmission link, save the extra UAV moving energy consumption and avoid the sensing task failure caused by UAV mobility.

DRLRR outperforms RLRR soon after the start of the relay process, which reduces the BER by 94.2% to 7×10^{-5}, decreases the outage probability by 93.8% to 2.0%, saves the robot energy consumption by 23.0% to 0.9 mJ, and increases the utility by 50.2% after 1500 time slots. The performance gain results from the three fully connected layer based DNNs that avoid choosing the robot mobility and relay policies related to transmission failure, and provide a more stable estimation of the expected long-term utility. The deep RL based robot relay is more effective than RLRR for the robot with sufficient computational resources such as RoboMaster EP, and suffers from performance degradation for the robot without enough resources to support deep learning.

The performance averaged over 100 runs and 1000 time slots in Figure 5 shows that our proposed relay schemes work well to resist a smart jammer whose maximum power changes from 10 to 50 mW. The BER, the outage probability and the robot energy consumption increase with the maximum jamming power, and the proposed schemes are more robust than QTC and STUD against strong jamming. For example, RLRR has BER lower than 2.6‰, outage probability lower than 43.6%, and energy consumption less than 1.1 mJ, if the maximum jamming power is lower than 50 mW. The performance gain of RLRR over QTC, including the BER, the outage probability and the energy consumption is greater than 64.4%, 18.6% and 38.5%, respectively, as the maximum jamming power changes between 10 and 50 mW. In addition, our scheme reduces the BER by 94.7%, decreases the outage probability by 53.7%, and saves the energy consumption by 30.2% compared with STUD even under strong jamming with power 50 mW. DRLRR further improves the performance in terms of the BER, the outage probability and the energy consumption by 62.4%, 62.5%, and 12.6%, respectively. As shown in Figure 5(d), the utility of RLRR

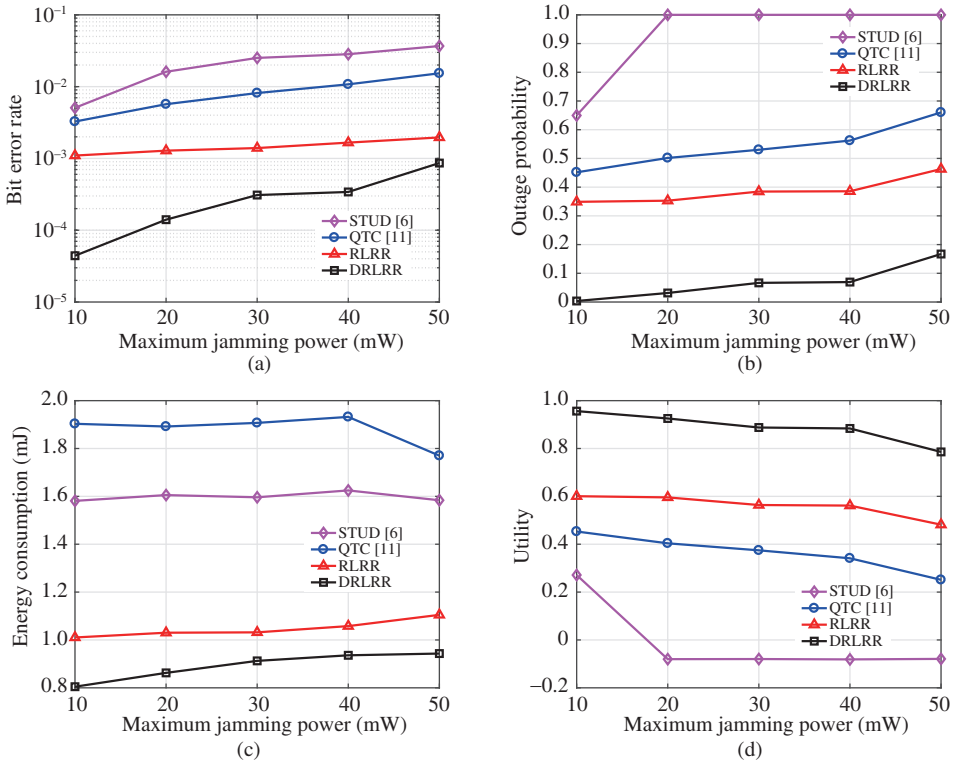

Figure 5 (Color online) Performance of the robot relay schemes averaged over 100 runs and 1000 time slots against a smart jammer whose maximum jamming power is given by $\{10, 20, 30, 40, 50\}$ mW. (a) BER; (b) outage probability; (c) energy consumption; (d) utility.

decreases with the maximum jamming power owing to the increase of the BER, the outage probability and the robot energy consumption. Its performance gain over QTC is higher than 28.1% even under severe jamming attacks, and DRLRR further improves the utility by 38.9%.

8 Conclusion

In this paper, we have proposed an RL-based robot relay scheme for UAVs in the presence of smart jamming attacks. By combining reinforcement learning with a tile coding technique, this scheme jointly optimizes the robot's moving distance and relay power with the goal of improving the UAV transmission quality and saving the robot energy. We have proposed a deep RL version with safe exploration that enables a robot with enough computing resources to choose the mobility and relay policy for a lower sample and computational complexity. The computational complexity is analyzed and the lower bounds containing the BER and the energy consumption are provided by deriving the Nash equilibrium of the robot relay game. Simulations have been performed based on a UAV-ground communication system against a smart jammer whose maximum jamming power is 10 mW, showing that the performance of the proposed RLRR and DRLRR exceeds the QTC [11] by reducing 67.5% and 98.1% BER, reducing 36.6% and 96.0% outage probability, and saving 31.4% and 47.2% robot energy after 1500 time slots.

In the future, we plan to improve the energy efficiency that consists of both the propulsion power and transmit power by incorporating the UAV trajectory optimization. More specifically, our scheme can choose to optimize the UAV trajectory or the robot relay policy that consists of the relay power and mobility to improve the anti-jamming performance and energy efficiency.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61971366, 61731012), Fundamental Research Funds for the Central Universities (Grant No. 20720200077), and Natural Science Foundation of Fujian Province of China (Grant No. 2020J01430).

References

1 You X, Wang C X, Huang J, et al. Towards 6G wireless communication networks: vision, enabling technologies, and new paradigm shifts. Sci China Inf Sci, 2021, 64: 110301

2 Sun X, Ng D W K, Ding Z, et al. Physical layer security in UAV systems: challenges and opportunities. IEEE Wireless Commun, 2019, 26: 40–47
3 Wu Q, Mei W, Zhang R. Safeguarding wireless network with UAVs: a physical layer security perspective. IEEE Wireless Commun, 2019, 26: 12–18
4 Yan Q, Zeng H, Jiang T, et al. Jamming resilient communication using MIMO interference cancellation. IEEE Trans Inform Forensic Secur, 2016, 11: 1486–1499
5 Luo S X, Zhang Z S, Wang S, et al. Network for hypersonic UCAV swarms. Sci China Inf Sci, 2020, 63: 140311
6 Lin Z, Lu X, Dai C, et al. Reinforcement learning based UAV trajectory and power control against jamming. In: Proceedings of International Conference on Machine Learning for Cyber Security, 2019. 336–347
7 Kim K H, Shin K G, Niculescu D. Mobile autonomous router system for dynamic (re)formation of wireless relay networks. IEEE Trans Mobile Comput, 2013, 12: 1828–1841
8 Liu D, Wang J, Xu K, et al. Task-driven relay assignment in distributed UAV communication networks. IEEE Trans Veh Technol, 2019, 68: 11003–11017
9 He X, Dai H, Ning P. Dynamic adaptive anti-jamming via controlled mobility. IEEE Trans Wireless Commun, 2014, 13: 4374–4388
10 Zeng Y, Zhang R, Lim T J. Throughput maximization for UAV-enabled mobile relaying systems. IEEE Trans Commun, 2016, 64: 4983–4996
11 Hu J, Zhang H, Song L, et al. Reinforcement learning for a cellular Internet of UAVs: protocol design, trajectory control, and resource management. IEEE Wireless Commun, 2020, 27: 116–123
12 Xiao L, Jiang D, Xu D, et al. Two-dimensional antijamming mobile communication based on reinforcement learning. IEEE Trans Veh Technol, 2018, 67: 9499–9512
13 Xiao L, Xie C, Min M, et al. User-centric view of unmanned aerial vehicle transmission against smart attacks. IEEE Trans Veh Technol, 2018, 67: 3420–3430
14 Sutton R S, Barto A G. Reinforcement Learning: An Introduction. Cambridge: MIT Press, 2018
15 Zeng Y, Xu X. Path design for cellular-connected UAV with reinforcement learning. In: Proceedings of IEEE Global Communications Conference, Waikoloa, 2019. 1–6
16 Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. Nature, 2015, 518: 529–533
17 Lyu J, Zeng Y, Zhang R, et al. Placement optimization of UAV-mounted mobile base stations. IEEE Commun Lett, 2017, 21: 604–607
18 Yanmaz E, Kuschnig R, Bettstetter C. Achieving air-ground communications in 802.11 networks with three-dimensional aerial mobility. In: Proceedings of IEEE INFOCOM, Turin, 2013. 120–124
19 Namvar N, Saad W, Bahadori N, et al. Jamming in the Internet of Things: a game-theoretic perspective. In: Proceedings of IEEE Global Communications Conference, Washington, 2016. 1–6
20 Guo J, Zhao N, Yu F R, et al. Exploiting adversarial jamming signals for energy harvesting in interference networks. IEEE Trans Wireless Commun, 2017, 16: 1267–1280
21 Jia L, Xu Y, Sun Y, et al. Stackelberg game approaches for anti-jamming defence in wireless networks. IEEE Wireless Commun, 2018, 25: 120–128
22 Gao N, Qin Z, Jing X, et al. Anti-intelligent UAV jamming strategy via deep Q-Networks. IEEE Trans Commun, 2020, 68: 569–581
23 Zhang L, Guan Z, Melodia T. United against the enemy: anti-jamming based on cross-layer cooperation in wireless networks. IEEE Trans Wireless Commun, 2016, 15: 5733–5747
24 Gu P, Hua C, Xu W, et al. Control channel anti-jamming in vehicular networks via cooperative relay beamforming. IEEE Internet Things J, 2020, 7: 5064–5077
25 Wang W, Tang J, Zhao N, et al. Joint precoding optimization for secure SWIPT in UAV-aided NOMA networks. IEEE Trans Commun, 2020, 68: 5028–5040
26 Pang X W, Tang J, Zhao N, et al. Energy-efficient design for mmWave-enabled NOMA-UAV networks. Sci China Inf Sci, 2021, 64: 140303
27 van Huynh N, Nguyen D N, Hoang D T, et al. "Jam Me If You Can:" defeating jammer with deep dueling neural network architecture and ambient backscattering augmented communications. IEEE J Sel Areas Commun, 2019, 37: 2603–2620
28 Xiao L, Jiang D, Chen Y, et al. Reinforcement-learning-based relay mobility and power allocation for underwater sensor networks against jamming. IEEE J Ocean Eng, 2020, 45: 1148–1156
29 Lu X, Xiao L, Dai C, et al. UAV-aided cellular communications with deep reinforcement learning against jamming. IEEE Wireless Commun, 2020, 27: 48–53
30 Avendi M R, Nguyen H H. Performance of selection combining for differential amplify-and-forward relaying over time-varying channels. IEEE Trans Wireless Commun, 2014, 13: 4156–4166
31 Hasselt H, Guez A, Silver D. Deep reinforcement learning with double Q-learning. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence, Phoenix, 2016. 2094–2100
32 Ou G, Murphey Y L. Multi-class pattern classification using neural networks. Pattern Recogn, 2007, 40: 4–18
33 Jin C, Allen-Zhu Z, Bubeck S, et al. Is Q-learning provably efficient? In: Proceedings of Conference on Advances in Neural Information Processing System, Montreal, 2018. 4868–4878
34 Cai Y, Cui F, Shi Q, et al. Dual-UAV-enabled secure communications: joint trajectory design and user scheduling. IEEE J Sel Areas Commun, 2018, 36: 1972–1985
35 Deyle T, Reynolds M. Surface based wireless power transmission and bidirectional communication for autonomous robot swarms. In: Proceedings of IEEE International Conference on Robotics and Automation, Pasadena, 2008. 1036–1041
36 Mozaffari M, Saad W, Bennis M, et al. Mobile unmanned aerial vehicles (UAVs) for energy-efficient Internet of Things communications. IEEE Trans Wireless Commun, 2017, 16: 7574–7589