

Intelligent Passive Eavesdropping in Massive MIMO-OFDM Systems via Reinforcement Learning

Jiawei Wang^{ID}, Peng Zhang^{ID}, Lan Tang^{ID}, *Member, IEEE*, Yechao Bai^{ID}, *Member, IEEE*,
and Lvxi Yang^{ID}, *Senior Member, IEEE*

Abstract—Massive multiple-input-multiple-output (MIMO) with narrow beam enhances the confidentiality of communication between base station and users, but also increases the difficulty for legal eavesdropping. In this letter, we study the passive eavesdropping scheme in the massive MIMO-OFDM systems by utilizing mobility of the monitor. Our objective is to maximize the average eavesdropping rate under the constraints of energy supply, moving direction and speed by jointly optimizing the receiving beamformers and moving trajectory. Due to the unknown environment knowledge and location of suspicious user, we propose the solution based on concatenated deep Q-network (DQN) to obtain the optimal solution. Simulation results verify the validity of the proposed method.

Index Terms—Reinforcement learning, passive eavesdropping, deep Q-network, massive MIMO-OFDM, hybrid beamforming.

I. INTRODUCTION

TO PREVENT crime or terrorist attacks, government agencies need to legally monitor suspicious communication links to detect abnormal behaviors.

So far, there are mainly two types of wireless eavesdropping: proactive eavesdropping and passive eavesdropping. In passive eavesdropping, the legitimate monitor remains radio silence and simply listens to the suspicious links. While proactive eavesdropping requires the monitor to deliberately broadcast jamming signals to reduce the data rate of suspicious users and improve the eavesdropping rate [1]. In a traditional three-node communication model, a proactive eavesdropping method with cognitive jamming was proposed [2]. Considering the proactive monitoring scenario under Rayleigh fading channel conditions, the average eavesdropping rate was improved in [3] through interference and power control. Reference [4] proposed a new proactive eavesdropping scheme with spoofing relay technology. In multiple-input-multiple-output (MIMO)

communication systems, [5] studied the achievable eavesdropping rates of multiple suspicious links with effective interference or relay by legal eavesdroppers.

All of above works assume that the monitor has prior knowledge of the suspicious communication link, which is impractical. Moreover, in 5G application scenarios, millimeter-wave (mm-wave) is always combined with massive MIMO to provide high beamforming gain. Massive MIMO system has the property of “secrecy for free” because the base station (BS) can significantly suppress eavesdroppers by narrow beamforming toward intended terminals [6]. When the monitor and suspicious user are not in the same beam coverage, both existing passive and proactive eavesdropping in [1]–[5] are difficult to achieve good performance. Therefore, in massive MIMO systems with narrow beam, it is critical to search for the suitable location for eavesdropping. When the channel and location of suspicious user are unknown and time-varying, it is time-consuming to find the optimal locations for eavesdropping in practical systems. To reduce the search time in the real world, one promising way is that the monitor learns the optimal moving strategy offline, with which the average eavesdropping rate is maximized, and then applies it online. Recently, applying reinforcement learning (RL) to optimize the moving trajectory of unmanned aerial vehicle (UAV) has been investigated in some literatures [7]–[10]. In these work, single-agent RL algorithms (Sarsa, Q-learning, et) and some multi-agent RL algorithms (distributed Q-learning, multi-agent DQN, et) were used to optimize the moving trajectory of UAV to maximize the performance metric of UAV communications. Specially, [10] maximized average secrecy rate via finding an optimal UAV trajectory against active eavesdropping with some RL algorithms (Expected SARSA, Double Q-learning, et). In most literatures about trajectory optimization of UAV, UAV acts as a service provider and its moving trajectory is optimized in a cooperative environment. Different from most of existing work, as a eavesdropper, the moving trajectory of UAV has to be optimized in an uncooperative environment. To the best of our knowledge, utilizing RL to find the suitable locations for eavesdropping has not been studied.

In this letter, to realize the legal eavesdropping, we propose an intelligent passive eavesdropping scheme in massive MIMO-OFDM systems. The target of this scheme is to maximize the discounted modified eavesdropping rate by optimizing the receiving beamformers and moving trajectory of the monitor. We assume that the monitor has no prior knowledge about the channel or location of the suspicious user. The difficulty is that in the real world, we don’t have sufficient training

Manuscript received February 24, 2022; accepted March 21, 2022. Date of publication March 30, 2022; date of current version June 10, 2022. This work was supported in part by the National Key R&D Program of China under Grant 2019YFE0113200, and in part by the National Natural Science Foundation of China under Grant U1936201, Grant 62072229, Grant 62071220, and Grant 61976113. The associate editor coordinating the review of this article and approving it for publication was M. Derakhshani. (Corresponding author: Lan Tang.)

Jiawei Wang, Lan Tang, and Yechao Bai are with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210093, China (e-mail: tanglan@nju.edu.cn).

Peng Zhang is with the Southwest Regional Air Traffic Management Bureau, CAAC, Chengdu 610200, China.

Lvxi Yang is with the School of Information Science and Engineering, Southeast University, Nanjing 210096, China.

Digital Object Identifier 10.1109/LWC.2022.3163268

time to find the optimal trajectory in an uncooperative environment. Moreover, the optimal trajectory is related to the location of the suspicious user. It is impractical to retrain the policy when the location of the suspicious user changes. Therefore, in the training phase, we assume that the user's location changes randomly in a given area so that the training can be implemented in a simulator without knowing the user's location. Then, we test the performance of the algorithm in the case of fixed suspicious users location to simulate a real scenario of eavesdropping. A concatenated DQN scheme is designed to decide the optimal receiving beamformers, moving direction and speed of the monitor. To make the receiving beam align with the transmitting beam, and suppress interferences from nearby BSs at the same time, the first DQN and minimum mean squared error (MMSE) receiver are used to determine the optimal analog and digital beamformers, respectively. With optimized beamformers, the second DQN is employed to get the optimal moving direction and speed. The simulation results show that the monitor, which trained by our proposed method, can find appropriate monitoring locations quickly in a real environment.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. System Model

We consider a legitimate surveillance system in multi-cell environment, where a removable monitor E aims to eavesdrop the communication information transmitted to the suspicious user D by the BS B, which operates in the mm-wave band. It is assumed that the BS is equipped with N_{rf} radio frequency (RF) chains connected to N_t transmit antennas ($N_{rf} < N_t$), and the terminal has only one RF chain connected with $N_{U,r}$ antennas. Because the BS uses massive antenna array, the beam directed to the suspicious user is narrow. When the monitor is not located in the beam coverage, passive eavesdropping becomes difficult. On the other hand, we assume that the monitor neither knows the location of the suspicious user nor channel state information (CSI) of B-D link so that it is difficult to implement proactive eavesdropping.

In this scenario, we will consider the eavesdropping scheme utilizing smart mobile device which can find the optimal monitoring location by learning during the movement. When the monitor is outside the narrow beam coverage, the monitoring process consists of two stages: in the first stage, the monitor searches for the suitable location to demodulate downlink control information (DCI) in control channel since reference signals (RS) and modulation and coding scheme (MCS) for demodulating DCI can be known in advance in practice [11]. During the movement, the monitor continuously utilizes RS to estimate channel and calculate signal to interference plus noise ratio (SINR), which guides the monitor to approach the desired beam coverage. If DCI is demodulated successfully, the monitor has entered the desired narrow beam coverage. Then, in the following stage, the monitor will utilize RS and MCS in DCI to estimate channel and demodulate the data information in physical downlink shared channel (PDSCH). Assume that the transmission rate between the BS B and user D is R_0 , which can be calculated according to MCS in DCI.

In the second phase, the monitor will search for the appropriate locations where the channel capacity of B-E link is greater than R_0 . When the monitor moves to these locations, the passive eavesdropping is achieved. In the two stages, we use the same learning algorithm to find the optimal monitoring location. Therefore, in the later analysis, we focus on the search process in the second stage with known DCI.

Under the hardware cost constraints of massive MIMO, hybrid beamforming is adopted to reduce the hardware cost [12]. To avoid interference, different users are allocated different subcarriers. Assuming that the subcarrier set $\mathcal{A}$ is allocated to the suspicious user D. The BS B first uses a digital precoder $\mathbf{f}_{bb_k}(t) \in \mathbb{C}^{N_{rf} \times 1} (k \in \mathcal{A})$ to precode the data symbol $s_k(t)$, and then an analog precoder $\mathbf{F}_{rf}(t) \in \mathbb{C}^{N_t \times N_{rf}}$ is employed after inverse fast fourier transform (IFFT) to generate the transmit signals.

The monitor is equipped with $N_{E,rf}$ RF chains connected to $N_{E,r}$ receiving antennas, and adopt hybrid beamforming to monitor data in $\mathcal{A}$. Because of the densely deployed mm-Wave BSs, the monitor is interfered by other surrounding BSs. At the time t , the received signals on the k th subcarrier can be expressed as

$$\mathbf{y}_{E,k}(t) = \mathbf{H}_{E,k}(t)\mathbf{f}_k(t)s_k(t) + \mathbf{I}_{E,k}(t) + \mathbf{n}_{E,k} \quad (1)$$

where $\mathbf{H}_{E,k}(t) \in \mathbb{C}^{N_{E,r} \times N_t}$ represents the channel matrix between E and B on the k th subcarrier, $\mathbf{f}_k(t) = \mathbf{F}_{rf}(t)\mathbf{f}_{bb_k}(t)$, $\mathbf{I}_{E,k}(t)$ is the interference from the surrounding BSs on the k th subcarrier and $\mathbf{n}_{E,k} \sim \mathcal{CN}(0, \sigma^2 \mathbf{I})$. In a geometric channel model [12], mm-wave channel for the k th subcarrier is modeled as

$$\mathbf{H}_{E,k}(t) = P_L \sum_{l=1}^{L_m} \alpha_l \mathbf{a}_r(\phi_l^r) \mathbf{a}_t^H(\phi_l^t) e^{-j2\pi l \frac{k}{K}} \quad (2)$$

where P_L is path loss inversely proportional to the transmission distance, L_m is the number of scatters, K is the number of subcarriers, $\alpha_l \sim \mathcal{CN}(0, \frac{N_t N_{E,r}}{L_m})$ is the small-scaled fading coefficient, ϕ_l^r and ϕ_l^t are the angles of arrival and departure for the l th path and $\mathbf{a}_r(\cdot)$ and $\mathbf{a}_t(\cdot)$ are the receive and transmit steering vectors, respectively.

The received signals $\{\mathbf{y}_{E,k}(t)\}_{k=1}^K$ are passed through a common analog combiner $\mathbf{W}_{E,rf}(t) \in \mathbb{C}^{N_{E,rf} \times N_{E,r}}$. As in [13], $\mathbf{W}_{E,rf}(t)$ is selected from a predefined discrete codebook $\mathcal{C}_E$. The process can be realized by deep reinforcement learning in dynamic environment as [14]. Then the cyclic prefix is removed and $N_{E,rf}$ K -point fast fourier transforms (FFTs) are used to demodulate signals. Finally, $|\mathcal{A}|$ low-dimensional digital combiners $\{\mathbf{w}_{E,bb_k}(t), k \in \mathcal{A}\}$ are used to combine signals in $\mathcal{A}$, where $|\mathcal{A}|$ denotes the size of $\mathcal{A}$. Therefore, the output of the k th combiner is

$$\begin{aligned} \tilde{s}_{E,k}(t) &= \mathbf{w}_{E,bb_k}(t) \mathbf{W}_{E,rf}(t) \mathbf{H}_{E,k}(t) \mathbf{f}_k(t) s_k(t) \\ &\quad + \mathbf{w}_{E,bb_k}(t) \mathbf{W}_{E,rf}(t) (\mathbf{I}_{E,k}(t) + \mathbf{n}_{E,k}). \end{aligned} \quad (3)$$

We assume that the monitor can estimate the equivalent channel $\mathbf{H}_{E,k}(t)\mathbf{f}_k(t)$ based on known RS. Meanwhile, the interference from neighbor BSs can be measured with reserved CSI-IM (interference measurement) resources [11]. At time t , monitor can use these channel information to design the

receiving beamformer and calculate SINR $\gamma_k(t)$ on the k th subcarrier

$$\gamma_k(t) = \frac{|\mathbf{w}_{E,bb_k}(t)\mathbf{W}_{E,rf}(t)\mathbf{H}_{E,k}(t)\mathbf{f}_k(t)|^2 p_k}{\mathbf{w}_{E,bb_k}(t)\mathbf{W}_{E,rf}(t)(\mathbf{C}_k(t) + \mathbf{I}\sigma^2)\mathbf{W}_{E,rf}^H(t)\mathbf{w}_{E,bb_k}^H(t)} \quad (4)$$

where the covariance matrix of $\mathbf{I}_{E,k}(t)$ is $\mathbf{C}_k(t)$ which can be estimated with measured interference, p_k is the transmission power of the BS B on the k th subcarrier. Based on (4), the channel capacity of the wiretap link at the time slot t is given by $R(t) = \sum_{k \in \mathcal{A}} \log(1 + \gamma_k(t))$.

B. Problem Formulation

Let $E(t)$ and $B(t)$ denote the battery energy and consumed power at time t , respectively. $B(t)$ includes the power used for moving (or hovering) and monitoring, denoted by $B_v(t)$ and $B_r(t)$, respectively. We assume that $B_v(t) = b_0$ when $v(t) = 0$, otherwise, $B_v(t) = b_1$, where $v(t)$ is the speed of the monitor at time t and $b_0 < b_1$. That means, the monitor will consume less energy when it stops to listen. The battery energy changes according to $E(t+1) = \max(0, E(t) - B(t)T)$, where T denotes the duration of a time slot. Due to the limited energy supply, in each time slot, the monitor needs to determine whether to move for the better monitoring location, which will consume additional energy for moving, or stop to listen. Meanwhile, it should be ensured that the monitor, such as UAV, has enough energy to return back to base. According to [3], the eavesdropping rate r'_t is R_0 if $R(t) \geq R_0$, otherwise, $r'_t = 0$. To get the monitor into the learning process as soon as possible in the initial stage, the eavesdropping rate is modified as $r_t = R_0$ when $R(t) \geq R_0$, otherwise, $r_t = R(t)$.

We define $(x(t), y(t))$ and $\theta(t)$ as the location and moving direction of the monitor at time t , respectively. When the monitor moves at a constant speed, $v(t) = v$. Otherwise, $v(t) = 0$. Our target is to maximize the expected discounted sum of modified eavesdropping rate under the constraint of energy consumption and moving speed. The joint optimization problem of receiving beamformers and moving direction and speed of the monitor can be expressed as

$$\max_{\{\theta(i), v(i), \mathbf{w}_{E,bb_k}(i), \mathbf{W}_{E,rf}(i)\}} \mathbb{E} \left(\sum_{i=0}^{\infty} \alpha^i r_{t+i} \right) \quad (5)$$

$$\text{s.t. } (B_v(i) + B_r(i))T + \frac{b_1 d(i+1)}{v} \leq E(i), \quad (5a)$$

$$\theta(i) \in \mathcal{C}_\theta, v(i) \in \{v, 0\}, \quad (5b)$$

$$\mathbf{W}_{E,rf}(i) \in \mathcal{C}_E. \quad (5c)$$

where $\alpha \in [0, 1]$ is the discount factor, $d(i+1) = \sqrt{x^2(i+1) + y^2(i+1)}$ represents the distance between the monitor and charging base at time $i+1$, θ denotes the angle between the moving direction of monitor and its head orientation, $x(i+1) = x(i) + v(i)T \cos(\theta(i))$, $y(i+1) = y(i) + v(i)T \sin(\theta(i))$, $\mathcal{C}_\theta$ is the set of the moving direction. The first part of (5a) denotes the energy required for the monitor to move and eavesdrop, and the second part is only the energy required to return back without monitoring. The constraint (5a) guarantees that the monitor has enough energy to return to the charging base.

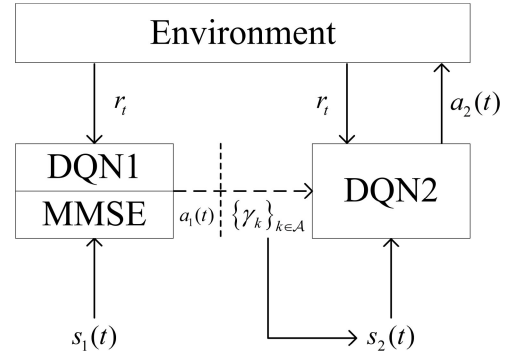


Fig. 1. Monitor receiver based on concatenated DQNs.

III. JOINT OPTIMIZATION OF RECEIVING BEAMFORMER AND MOVING TRAJECTORY

Obviously, the optimal solution to maximize the average modified eavesdropping rate is to find the suitable eavesdropping location and stop to eavesdrop. The difficult is that, we don't have sufficient training time to find the optimal trajectory in an uncooperative environment, and when the suspicious user moves, the trajectory may change. Therefore, the suspicious user's location is changed randomly in the training phase to improve robustness of the policy learned by monitor. The problem in (5) can be modeled as a markov decision process (MDP). When the system dynamic is unknown, we apply the RL to solve it.

In RL, we set the system state as $s(t) = \{\{\mathbf{H}_{E,k}(t)\mathbf{f}_k(t), \mathbf{C}_k(t)\}_{k \in \mathcal{A}}, E(t), x(t), y(t)\}$, and the action $a(t) = \{\theta(t), v(t), \{\mathbf{w}_{E,bb_k}(t)\}_{k \in \mathcal{A}}, \mathbf{W}_{E,rf}(t)\}$. When the monitor moves, the current position and energy will change, which leads to changes in channel and interference level. However, it is difficult to achieve good performance by applying DQN directly to find the optimal mapping between $s(t)$ and $a(t)$ due to the large action space [15]. To reduce the dimension of the state/action spaces, a monitor receiver design scheme based on concatenated DQNs is proposed. Since the monitor only moves a very short distance within a time slot, channels and covariance matrix of interference can be regarded as approximately invariant when the monitor moves from $\{x(t), y(t)\}$ to $\{x(t+1), y(t+1)\}$. Under this assumption, the action of the monitor can be determined sequentially. Specifically, the monitor first determine $a_1(t) = \{\mathbf{W}_{E,rf}(t)\}$. Then, $a_2(t) = \{\theta(t), v(t)\}$ is determined based on $s_2(t) = \{x(t), y(t), E(t), \{\gamma_k(t)\}_{k \in \mathcal{A}}\}$ as shown in Fig. 1. The detailed process of DQN refers to [16]. DQN1 and MMSE module are combined to generate the analog and digital combination matrices $\mathbf{W}_{E,rf}(t)$ and $\{\mathbf{w}_{E,bb_k}(t)\}_{k \in \mathcal{A}}$ based on $s_1(t)$. To improve SINR $\gamma_k(t)$, we first adopt the MMSE principle $\min_{\mathbf{w}_{E,bb_k}(t)} \mathbb{E}\{|\mathbf{w}_{E,bb_k}(t)\mathbf{W}_{E,rf}(t)\mathbf{y}_{E,k}(t) - s_k(t)|^2\}$ to calculate $\mathbf{w}_{E,bb_k}(t)$.

For given $\mathbf{W}_{E,rf}(t)$ and $\{\mathbf{C}_k(t)\}_{k \in \mathcal{A}}$, the MMSE solution is

$$\mathbf{w}_{E,bb_k}(t) = p_k \mathbf{X}_k^H(t) (p_k \mathbf{X}_k(t) \mathbf{X}_k^H(t) + \mathbf{W}_{E,rf}(t)(\mathbf{C}_k(t) + \sigma^2 \mathbf{I}) \mathbf{W}_{E,rf}^H(t))^{-1}. \quad (6)$$

where $\mathbf{X}_k(t) = \mathbf{W}_{E,rf}(t)\mathbf{H}_{E,k}(t)\mathbf{f}_k(t)$. Then we substitute (6) into (4) to get the expression of $R(t)$. And then DQN1 is utilized to select the optimal $\mathbf{W}_{E,rf}(t)$ from $\mathcal{C}_E$ based on $s_1(t)$ to maximize the reward function in (5).

With the obtained beamformers, we can calculate $\{\gamma_k(t)\}_{k \in \mathcal{A}}$ and make them as a part of the inputs of DQN2. The following DQN2 is used to determine the moving direction $\theta(t)$ and speed $v(t)$ based on $s_2(t) = \{x(t), y(t), E(t), \{\gamma_k(t)\}_{k \in \mathcal{A}}\}$. After obtaining $\{\mathbf{W}_{E,rf}(t), \mathbf{w}_{E,bk \in \mathcal{A}}(t)\}$, the problem in (5) can be simplified as

$$\max_{\{\theta(i), v(i)\}} \mathbb{E} \left(\sum_{i=0}^{\infty} \alpha^i r_{t+i} \right) \quad \text{s.t. (5a), (5b)}. \quad (7)$$

In (7), the system state and action are reduced to $s_2(t) = \{x(t), y(t), E(t), \{\gamma_k(t)\}_{k \in \mathcal{A}}\}$ and $a_2(t) = \{\theta(t), v(t)\}$. We can see that when serial scheme is considered, DQN1 and DQN2 have smaller action/state dimensions. In each episode, monitor leaves from the base and aims to learn a path to maximize the objective function and when the constraint (5a) is not satisfied, the episode is terminal.

According to [17], the time complexity of DQN is proportional to $|\mathcal{S}||\mathcal{A}|$, where $|\mathcal{S}|$ and $|\mathcal{A}|$ are sizes of state and action space, respectively. If we use single DQN scheme, the size of action space is $|\mathcal{A}_1| \times |\mathcal{A}_2|$. In the concatenated DQNs scheme, the size of action space is $|\mathcal{A}_1| + |\mathcal{A}_2|$, which is much smaller than $|\mathcal{A}_1| \times |\mathcal{A}_2|$ in our simulation settings.

IV. SIMULATION RESULTS

In this section, we will provide some numerical simulation results to show the performance of the proposed intelligent passive eavesdropping scheme. We set $N_t = 64$, $N_{rf} = 4$, $N_{E,r} = 16$, $N_{E,rf} = 4$, $N_r = 16$, the number of subcarriers is 16 and carrier frequency is 26GHz. We describe the location by Cartesian coordinates in meters. We consider a 150m-by-150m square area, where the BS is located at (0, 0), and the two interference BSs are located at (30m, 30m) and (20m, 50m), respectively. The monitor always travels from an initial location (70m, -40m). The location of the suspicious user is random in a circle area centered on (50m, 0) and 30m as the radius in each episode. The signal-to-noise ratio $\text{SNR} = \frac{p_k}{\sigma^2}$ is 20 dB. We suppose that in the B-D link, the ϕ_l^r and ϕ_l^t are uniformly distributed in $[0, 2\pi]$, the number of scatters is $L_m = 16$ and $\mathbf{f}_k(t)$, $\mathcal{C}_E$ are generated as in [13]. Based on the parameters above, the transmission rate R_0 in the B-D link is calculated as the sum of capacity of all subcarriers in $\mathcal{A}$.

We set $\theta(t) \in \{\frac{\pi}{4}i\}, i = 0, \dots, 7$ and the maximum number of episodes is 1000. In both DQN1 and DQN2, we adopt the two-layer Rectified Linear Unit (ReLU) Multi-Layer Perceptron (MLP) with 48 units per layer. The replay buffer size is 10^5 , the batch size is 256 and the learning rate is set to 0.001.

Fig. 2 shows the learning curves of the average modified eavesdropping rate $\mathbb{E}\{\frac{1}{T_m} \sum_{t=1}^{T_m} r_t\}$ denoted by solid lines and the average eavesdropping rate $\mathbb{E}\{\frac{1}{T_m} \sum_{t=1}^{T_m} r'_t\}$ denoted by dashed lines, where T_m represents the time interval from leaving base to returning back. The initial energy of monitor

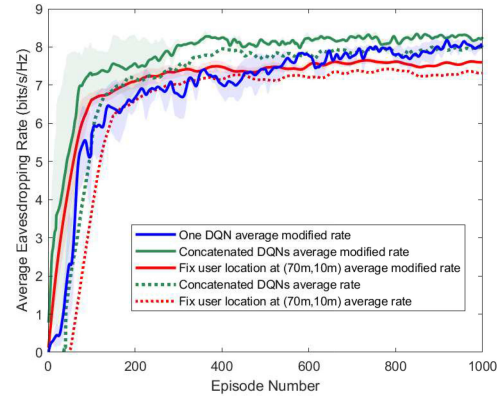


Fig. 2. The learning curves of average eavesdropping rate in different schemes.

is set to $E(0) = 800$ J. Because the location and channel of the suspicious user are random in every episode, after each episode of training is completed, the experiment is repeated multiple times to obtain the average value based on the currently trained strategy. We use the average (modified) eavesdropping rate instead of the objective function as a performance measure because the average eavesdropping rate has a clearer physical meaning. The solid green and blue lines denote the learning curves of the average modified eavesdropping rate in concatenated DQNs and single DQN scheme, respectively. It can be observed that the proposed scheme has a faster convergence speed and higher eavesdropping rate than single DQN scheme due to the reduced action space. The red lines are the learning curves of the average eavesdropping rate and average modified eavesdropping rate when the location of the suspicious user is fixed at (70m, 10m). We can see that red learning curves have smaller fluctuations than green ones. If the location of suspicious user remains unchanged so that we have enough learning time in a non-cooperative environment, the monitor will learn an optimal eavesdropping policy for this fixed user after about 300 episodes, as shown by the red solid line. If the location of suspicious user is unknown and we cannot get enough training time in practice, we will randomly change the user's position during the training phase in the simulator (green learning curves), and then apply the steady policy in the real system. In the initial stage, the average eavesdropping rate is smaller than the average modified rate because many transient rewards are 0. However, the gap between the average rate and the average modified one is vanishing as the number of learning episodes increases. Therefore, the modified reward can better guide the monitor in the initial stage, and results in the steady-state performance similar to the average eavesdropping rate.

Fig. 3 compares the average modified eavesdropping rates under different moving and monitoring policies with varying initial energy. We consider three specific user locations: (30m, -10m) (triangle), (50m, 0) (circle) and (70m, 10m) (square). When the user D is in these locations, the transmission rate R_0 is 10.1940bits/s/Hz, 8.7010bits/s/Hz and 8.4316bits/s/Hz, respectively. When the location of suspicious user is fixed during the training phase, the optimal policy learned by monitor

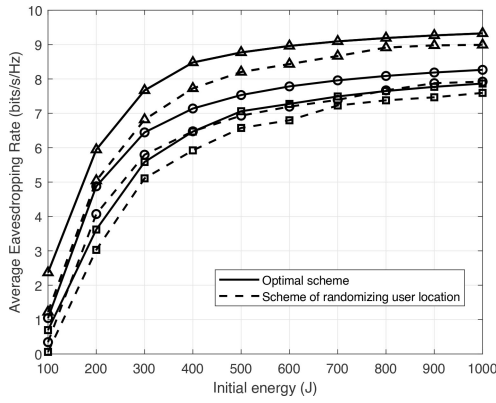


Fig. 3. The average eavesdropping rate of monitor vs. different initial energy.

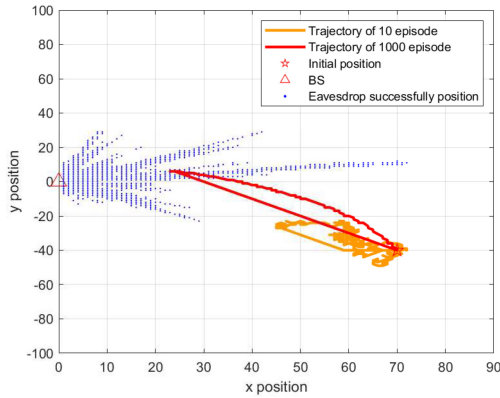


Fig. 4. The trajectory of monitor after different episodes.

based on concatenated DQNs is that UAV flies to an optimal location and eavesdrops until it returns back. The solid lines display the average eavesdropping rate under such optimal policy. Considering monitor has no prior knowledge of the optimal eavesdropping location in practical non-cooperative environment, we use the steady policy obtained by randomizing user's location to guide the monitor to move and eavesdrop and the corresponding eavesdropping rates are shown by dashed lines. We find that the UAV using "scheme of randomizing user location" can still find its ideal eavesdropping location and achieve performance approximated to the optimal policy. In the steady-state strategy, the monitor stops at the place where it can successfully eavesdrop most of the time. Therefore, the larger the R_0 is, the higher the average modified eavesdropping rate is.

Fig. 4 plots the moving trajectory learned by monitor after different episodes using concatenated DQNs scheme. We fix user's location at (70m, 10m) to show the moving trajectory of UAV with the policy learned by "scheme of randomizing user location". Blue points mark all the locations where monitor can eavesdrop successfully and orange and red lines denote the trajectory under the policy learned after 10 and 1000 episodes, respectively. It can be seen that the monitor will learn a proper policy, i.e., hovering at one certain position to eavesdrop, after enough episodes.

V. CONCLUSION

In this letter, an intelligent passive eavesdropping scheme in the massive MIMO-OFDM communication system is investigated to maximize the discounted eavesdropping rate without exact channel information or location of the suspicious user. A concatenated DQN scheme is designed to decide the optimal receiving beamformers and moving direction and speed of the monitor. The first DQN and MMSE receiver are used to determine the optimal analog and digital beamformers. With optimized beamformers, the second DQN is employed to get the optimal moving trajectory. The simulation results show that the proposed method can find appropriate monitoring locations quickly after about 1000 episodes.

REFERENCES

- [1] J. Moon, H. Lee, C. Song, S. Lee, and I. Lee, "Proactive eavesdropping with full-duplex relay and cooperative jamming," *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6707–6719, Oct. 2018.
- [2] J. Xu, L. Duan, and R. Zhang, "Surveillance and intervention of infrastructure-free mobile communications: A new wireless security paradigm," *IEEE Wireless Commun.*, vol. 24, no. 4, pp. 152–159, Aug. 2017.
- [3] J. Xu, L. Duan, and R. Zhang, "Proactive eavesdropping via jamming for rate maximization over Rayleigh fading channels," *IEEE Wireless Commun. Lett.*, vol. 5, no. 1, pp. 80–83, Feb. 2016.
- [4] Y. Zeng and R. Zhang, "Wireless information surveillance via proactive eavesdropping with spoofing relay," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 8, pp. 1449–1461, Dec. 2016.
- [5] B. Li, Y. Yao, H. Chen, Y. Li, and S. Huang, "Wireless information surveillance and intervention over multiple suspicious links," *IEEE Signal Process. Lett.*, vol. 25, no. 8, pp. 1131–1135, Aug. 2018.
- [6] A. Bereyhi, S. Asaad, R. R. Muller, R. F. Schaefer, and A. M. Rabiei, "On robustness of massive MIMO systems against passive eavesdropping under antenna selection," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2018, pp. 1–7.
- [7] S. Yin, S. Zhao, Y. Zhao, and F. R. Yu, "Intelligent trajectory design in UAV-aided communications with reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8227–8231, Aug. 2019.
- [8] B. Khamidehi and E. S. Sousa, "Reinforcement learning-based trajectory design for the aerial base stations," in *Proc. IEEE 30th Annu. Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2019, pp. 1–6.
- [9] F. Wu, H. Zhang, J. Wu, and L. Song, "Cellular UAV-to-device communications: Trajectory design and mode selection by multi-agent deep reinforcement learning," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4175–4189, Jul. 2020.
- [10] M. T. Mamaghani and Y. Hong, "Intelligent trajectory design for secure full-duplex MIMO-UAV relaying against active eavesdroppers: A model-free reinforcement learning approach," *IEEE Access*, vol. 9, pp. 4447–4465, 2021.
- [11] E. Dahlman, S. Parkvall, and J. Skold, *5G NR: The Next Generation Wireless Access Technology*. Amsterdam, The Netherlands: Academic, 2020.
- [12] F. Sohrabi and W. Yu, "Hybrid analog and digital beamforming for mmWave OFDM large-scale antenna arrays," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1432–1443, Jul. 2017.
- [13] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.
- [14] J. Zhang, Y. Huang, J. Wang, and X. You, "Intelligent beam training for millimeter-wave communications via deep reinforcement learning," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2019, pp. 1–7.
- [15] Z. Zhao, Y. Liang, and X. Jin, "Handling large-scale action space in deep Q network," in *Proc. Int. Conf. Artif. Intell. Big Data (ICAIBD)*, 2018, pp. 93–96.
- [16] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [17] G. Li, Y. Wei, Y. Chi, Y. Gu, and Y. Chen, "Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction," *IEEE Trans. Inf. Theory*, vol. 68, no. 1, pp. 448–473, Jan. 2022.