DeepAntiJam: Stackelberg Game-Oriented Secure Transmission via Deep Reinforcement Learning

Jianzhong Lu, Dongxuan He, and Zhaocheng Wang, *Fellow, IEEE*

***Abstract*— In this letter, we present a novel deep reinforcement learning-assisted anti-jamming transmission scheme (DeepAntiJam) to guarantee reliable and creditable transmission in the presence of one smart jammer and multiple eavesdroppers. Specifically, we formulate secure transmission as a Stackelberg game in which the jammer, as the leader, adaptively adjusts its jamming power while the transmitter, as the follower, selects its transmit power and secrecy rate accordingly. Furthermore, the existence and uniqueness of Stackelberg equilibrium are proved. To achieve the Stackelberg equilibrium when the prior knowledge of jammer is unknown, DeepAntiJam is proposed to improve the secrecy throughput, where a solver neural network is utilized to determine the optimal transmission parameter according to the transmission strategy obtained by deep reinforcement learning. Moreover, transfer learning is introduced into the initialization of DeepAntiJam to avoid unnecessary initial random exploration. Simulation results validate that DeepAntiJam can enhance secrecy throughput significantly under the coexistence of smart jammer.**

***Index Terms*— Reinforcement learning, smart jammer, multiple eavesdroppers, Stackelberg game, transfer learning.**

I. Introduction

DUE to its shortcomings in secret key generation, distribution and management, conventional cryptography-based protocols would fail to satisfy the escalating secrecy requirement of future wireless communications. As an alternative, physical layer security has shown to be effective in ensuring information confidentiality by exploiting the randomness of wireless channels, which has attracted much attention from both academia and industry [1].

Recently, game theory based solutions are popularly used in physical layer security to overcome the serious threat from reactive jammers [2], [3]. In particular, as a method focusing on hierarchical decision-making issues, Stackerlberg game is effective to analyze the mutual interactions between transmitters and jammers [4]. For example, by modeling the legitimate user as leader and the jammer as follower, Stackelberg game was utilized to control the transmit power [5]. Moreover, the observation inaccuracy of jammer in such model was further investigated in [6]. In addition, the Stackelberg approach was also used to obtain the optimal secrecy code rate and guard zone of eavesdroppers to guarantee secrecy performance [3].

Benefit from artificial intelligence, smart jammers are capable of inferring the transmission policy of transmitter and adjusting their strategy accordingly, which have become a tremendous challenge to secure wireless communications [7]. To address this issue, reinforcement learning (RL) is introduced to generate adaptive anti-jamming strategies [8]. For instance, a hierarchical discrete power control algorithm (HPCA) was proposed to obtain the Stackelberg equilibrium (SE) under the coexistence of Q-learning based jammer [9]. In [10], a fast Q-based power allocation scheme combining hotbooting technique and Dyna structure was investigated in a nonorthogonal multiple access system, which improves the anti-jamming performance significantly. However, the above works ignored the transmission confidentiality during anti-jamming, especially under the coexistence of multiple eavesdroppers.

In this letter, a deep RL-based transmission scheme is designed to ensure reliable and creditable transmission in the presence of one learning-assisted jammer and multiple eavesdroppers. Our main contributions are summarized as follows: 1) To resist the smart jammer and multiple passive eavesdroppers simultaneously, Stackelberg game is utilized to analyze the dynamic interactions between transmitter and jammer considering connection outage probability (COP) and secrecy outage probability (SOP), and the existence and uniqueness of the SE are mathematically proved. 2) A deep RL-based anti-jamming transmission scheme, called DeepAntiJam, is proposed to generate the optimal transmission strategy without any prior knowledge of jamming model, where a solver neural network (NN) is proposed to determine the transmission parameter based on the strategy obtained by deep RL. Specifically, transfer learning is introduced into initialization, which simplifies the initial random exploration. In comparison to its conventional anti-jamming counterparts, our proposed DeepAntiJam could improve the secrecy throughput remarkably in the presence of one Q-learning based or deep RL-based jammer.

II. Problem Formulation

A. System Model

As shown in Fig. 1, we consider the wireless link from transmitter (Alice) to legitimate receiver (Bob) in the presence of one smart jammer and M passive eavesdroppers E_m $(1 \leq m \leq M)$. To ensure the transmission confidentiality, the Wyner's encoding scheme with rate (R_t, R_s) is adopted, where R_t and R_s denote the transmission rate and secrecy rate,

Manuscript received 7 May 2022; revised 5 June 2022; accepted 7 June 2022. Date of publication 10 June 2022; date of current version 11 September 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62101306, and in part by the Postdoctoral Science Foundation of China under Grant 2020M670332. The associate editor coordinating the review of this letter and approving it for publication was C. Huang. *(Corresponding authors: Dongxuan He; Zhaocheng Wang.)*

Jianzhong Lu and Zhaocheng Wang are with the Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China, and also with the Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China (e-mail: ljz19@mails.tsinghua.edu.cn; zcwang@tsinghua.edu.cn).

Dongxuan He is with the Beijing National Research Center for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: dongxuan_he@mail.tsinghua.edu.cn).

Digital Object Identifier 10.1109/LCOMM.2022.3182001

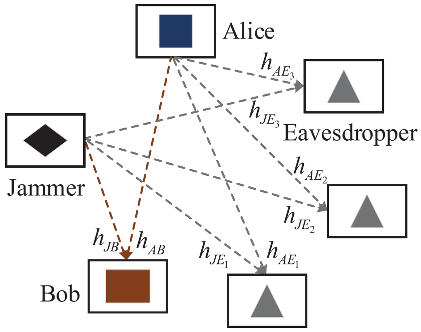

Fig. 1. Illustration of the secure transmission system.

respectively. Alice adaptively changes the secrecy rate R_s as well as its transmit power P_A to resist jamming. Meanwhile, the smart jammer, which is capable of learning the transmission strategy of Alice, adjusts its jamming power P_J flexibly.

All channels are assumed to experience independent and identically distributed (i.i.d) Rayleigh fading while only the statistical channel state information (CSI) is available at Alice and jammer. The instantaneous CSI of links Alice-Bob, Alice-E_m, Jammer-Bob, and Jammer-E_m are denoted by h_{AB}, h_{AE_m}, h_{JB}, and h_{JE_m}, respectively, which are complex Gaussian random variables with zero means and variances σ_{AB}^2, $\sigma_{AE_m}^2$, σ_{JB}^2, and $\sigma_{JE_m}^2$, respectively. Moreover, the noise powers at Bob and E_m are σ_B^2 and $\sigma_{E_m}^2$. Thus, the signal-to-interference-plus-noise ratio (SINR) at Bob and E_m can be derived as $\gamma_B = \frac{P_A|h_{AB}|^2}{P_J|h_{JB}|^2+\sigma_B^2}$ and $\gamma_{E_m} = \frac{P_A|h_{AE_m}|^2}{P_J|h_{JE_m}|^2+\sigma_{E_m}^2}$, respectively.

Typically, a connection outage occurs if the capacity of Bob is less than the transmission rate, i.e., $C_B < R_t$, and a secrecy outage occurs when the capacity of eavesdroppers is larger than the redundancy rate against wiretapping, i.e., $C_E > R_t - R_s$ [11]. To depict the transmission reliability, COP can be presented as

$$\begin{aligned}
P_{co} &= \mathbf{P}\left\{C_B < R_t\right\} = \mathbf{P}\left\{\frac{P_A|h_{AB}|^2}{P_J|h_{JB}|^2+\sigma_B^2} < \gamma_1\right\},\\
&= \mathbf{E}_{|h_{JB}|^2}\left\{1-\exp\left(-\frac{\gamma_1(P_J|h_{JB}|^2+\sigma_B^2)}{P_A\sigma_{AB}^2}\right)\right\},\\
&= 1-\exp\left(-\frac{\gamma_1\sigma_B^2)}{P_A\sigma_{AB}^2}\right)\\
&\quad\times\int_0^\infty \exp\left(-\frac{\gamma_1 P_J|h_{JB}|^2}{P_A\sigma_{AB}^2}\right)\frac{\exp(-\frac{|h_{JB}|^2}{\sigma_{JB}^2})}{\sigma_{JB}^2}d|h_{JB}|^2,\\
&= 1-\frac{\exp\left(-\frac{\gamma_1\sigma_B^2}{P_A\sigma_{AB}^2}\right)P_A\sigma_{AB}^2}{P_A\sigma_{AB}^2+\gamma_1 P_J\sigma_{JB}^2},
\end{aligned}\tag{1}$$

where $\gamma_1 = 2^{R_t} - 1$.

The passive eavesdroppers are assumed to operate at the non-colluding mode and their performance is determined by the eavesdropper with the strongest received signal, where the SINR of eavesdroppers can be formulated as $\max_{1\le m\le M}\{\gamma_{E_m}\}$. Therefore, SOP can be presented as

$$\begin{aligned}
P_{so} &= \mathbf{P}\left\{C_E > R_t - R_s\right\} = \mathbf{P}\left\{\max_{1\le m\le M}\left\{\gamma_{E_m}\right\} > \gamma_2\right\},\\
&= 1-\prod_{m=1}^{M}\mathbf{P}\left\{\frac{P_A|h_{AE_m}|^2}{P_J|h_{JE_m}|^2+\sigma_{E_m}^2} \le \gamma_2\right\},\\
&= 1-\prod_{m=1}^{M}\Bigg[1-\exp\left(-\frac{\gamma_2\sigma_{E_m}^2}{P_A\sigma_{AE_m}^2}\right)\\
&\quad\times\int_0^\infty \exp\left(-\frac{\gamma_2 P_J t_m}{P_A\sigma_{AE_m}^2}\right)\frac{1}{\sigma_{JE_m}^2}\exp(-\frac{t_m}{\sigma_{JE_m}^2})dt_m\Bigg],\\
&= 1-\prod_{m=1}^{M}\left[1-\frac{\exp\left(-\frac{\gamma_2\sigma_{E_m}^2}{P_A\sigma_{AE_m}^2}\right)P_A\sigma_{AE_m}^2}{P_A\sigma_{AE_m}^2+\gamma_2 P_J\sigma_{JE_m}^2}\right],
\end{aligned}\tag{2}$$

where $\gamma_2 = 2^{R_t-R_s} - 1$ and $t_m = |h_{JE_m}|^2$.

Moreover, to facilitate the analysis of Stackelberg game, the monotonicity of P_{co} and P_{so} are presented in the following lemma.

Lemma 1: P_{co} is a decreasing function of P_A, and P_{so} is an increasing function of P_A and R_s.

Proof: Please refer to Appendix. ■

B. Stackelberg Game Model

To obtain the optimal transmission strategy against smart jamming, the interactions between Alice and smart jammer are formulated as a Stackelberg game, where the jammer acts as leader and Alice acts as follower [4]. Due to the adaptation of jamming policy, Alice cannot determine its optimal transmission strategy directly, which requires Alice to dynamically select its transmission parameters. Specifically, the smart jammer as the leader first imposes the interference strategy by elaborately selecting a jamming power from its power set $\mathcal{P}_J = \{P_1, P_2, \ldots, P_J^{max}\}$ to deteriorate the secure transmission probability P_s. Then, Alice accordingly updates its anti-jamming strategy with the purpose of maximizing the secrecy throughput T_s by fine-tuning its transmit power P_A, where $0 \le P_A \le P_A^{max}$, and selecting its secrecy rate from $\mathcal{R}_s = \{R_1, R_2, \ldots, R_s^{max}\}$. Here, T_s is defined as the average number of successfully transmitted secrecy bits per second per Hz, i.e., $T_s = (1 - P_{co})R_s$ [11]. To be specific, P_s is utilized to characterize the security and reliability performance simultaneously, given by $P_s = (1 - P_{co})(1 - P_{so})$. Thus, the utility functions of jammer and Alice are $U_J(P_A, R_s, P_J) = 1 - P_s$ and $U_A(P_A, R_s, P_J) = T_s$, respectively.

To avoid being detected by legitimate users directly, the jamming power of jammer is controlled under a threshold P_J^{max}. Therefore, the optimization problem of jammer can be expressed as

$$\max_{P_J}\ 1-(1-P_{co})(1-P_{so}) \quad s.t.\ P_J \le P_J^{max}. \tag{3}$$

Meanwhile, the optimization problem of Alice can be formulated as

$$\max_{R_s,P_A}(1-P_{co})R_s \quad s.t.\ P_{so} \le \varepsilon, \tag{4}$$

where ε is a certain threshold guaranteeing the SOP derived in (2).

C. Stackelberg Equilibrium

At the SE of our considered anti-jamming game described in Section II-B, the smart jammer properly selects its jamming strategy to maximize its expected utility U_J based on the estimated P_A and R_s. Meanwhile, based on the observed jamming power P_J, Alice determines its transmit power P_A and secrecy rate R_s to maximize its throughput T_s. As a result, the SE constituted by (R_s^*, P_A^*, P_J^*) can be expressed as

$$\begin{cases} P_J^* = \underset{P_J \in \mathcal{P}_J}{\arg\max}\ U_J\left(R_s, P_A, P_J\right), \\ R_s^*, P_A^* = \underset{0 \le P_A \le P_A^{max}, R_s \in \mathcal{R}_s}{\arg\max} U_A\left(R_s, P_A, P_J\right). \end{cases} \tag{5}$$

To prove the existence of equilibrium (5), the follower's problem (4) should be solved by two steps: 1) given R_s, maximize $1 - P_{co}$ over available P_A; 2) maximize T_s over the available R_s. Due to that P_{co} decreases w.r.t. P_A and P_{so} increases w.r.t. P_A, for a given R_s, the optimal $P_A^*(R_s)$ could be obtained by solving $P_{so} = \varepsilon$.

Furthermore, we maximize $T_s(R_s)$ over R_s with the help of following lemma.

Lemma 2: The secrecy throughput $T_s(R_s)$ is a quasi-concave function of R_s.

Proof: Take the first-order derivative of $T_s(R_s)$ w.r.t. R_s, we have

$$T_s' = \frac{\partial T_s(R_s)}{\partial R_s} = 1 - P_{co} - R_s \frac{\partial P_{co}}{\partial R_s}, \tag{6}$$

where

$$\frac{\partial P_{co}}{\partial R_s} = \frac{\partial P_{co}}{\partial P_A^*(R_s)} \frac{\partial P_A^*(R_s)}{\partial R_s}. \tag{7}$$

Define $F = P_{so} - \varepsilon$, the first-order derivative of $P_A^*(R_s)$ w.r.t. R_s can be obtained according to the derivative rule for implicit function, given by

$$\frac{\partial P_A^*(R_s)}{\partial R_s} = -\frac{\frac{\partial F}{\partial R_s}}{\frac{\partial F}{\partial P_A^*(R_s)}} = -\frac{\frac{\partial P_{so}}{\partial R_s}}{\frac{\partial P_{so}}{\partial P_A^*(R_s)}} < 0. \tag{8}$$

Based on (8), $P_A^*(R_s)$ decreases monotonically w.r.t. R_s. Hence, we have $\frac{\partial P_{co}}{\partial R_s} > 0$. Thus, $T_s'(0) > 0$ and $T_s'(\infty) < 0$ can be obtained. Accordingly, T_s is a quasi-concave function of R_s and there exists a unique R_s^* that maximizes T_s. ■

The aforementioned **Lemma 1** proves that the best-response strategy set of the follower is a singleton, which corresponds to the unique secrecy rate R_s and transmit power P_A. As a result, the existence of equilibrium in our considered scenario can be guaranteed [12].

Based on the analysis of SE, for a fixed jamming policy the optimal transmission strategy can be obtained numerically [3]. However, when the jamming policy is dynamic and unknown, it is still challenging to obtain the optimal transmission strategy, which should be further investigated.

III. DeepAntiJam Methodology

To update the transmission policy without any prior knowledge about the jamming model, DeepAntiJam relying on deep RL is proposed, which can accelerate the optimization of transmission policy and suppress smart jamming.

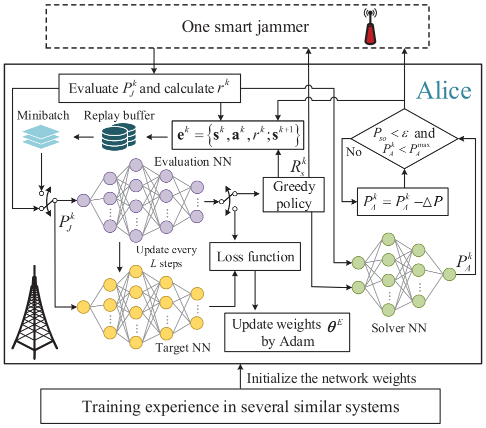
One smart jammer
Alice
Evaluate P_J^k and calculate r^k
Minibatch
Replay buffer
$\mathbf{e}^k = \{\mathbf{s}^k, \mathbf{a}^k, r^k; \mathbf{s}^{k+1}\}$
Evaluation NN
P_J^k
R_s^k
Greedy policy
$P_{so} < \varepsilon$ and $P_A^k < P_A^{max}$
No
$P_A^k = P_A^k - \Delta P$
Update every L steps
Loss function
P_A^k
Target NN
Update weights $\boldsymbol{\theta}^E$ by Adam
Solver NN
Initialize the network weights
Training experience in several similar systems

Fig. 2. Illustration of our proposed DeepAntiJam.

As illustrated in Fig. 2, Alice firstly observes the state $\mathbf{s}^k$ at time k, which consists of its estimated jamming power at time $k-1$, i.e., P_J^{k-1}. As a result, the state $\mathbf{s}^k$ can be expressed as $\mathbf{s}^k = \{P_J^{k-1}\}$. Next, a two-step actions selection is executed to generate the action of Alice, defined as $\mathbf{a}^k = \{R_s^k, P_A^k\}$, where the secrecy rate R_s^k is selected according to the Q-function values computed by the deep neural network and then the transmission power P_A^k is determined according to $P_{so} = \varepsilon$. Then, the utility denoted by r^k can be calculated based on (1). At time $k+1$, the experience $\mathbf{e}^k$ including $\{\mathbf{s}^k; \mathbf{a}^k; r^k; \mathbf{s}^{k+1}\}$ which records one interaction between jammer and Alice is stored in the replay buffer. Note that the Q-function denoted by $Q(\mathbf{s}, \mathbf{a}; \boldsymbol{\theta})$ corresponds to the long-term discounted reward when Alice takes action $\mathbf{a}$ at state $\mathbf{s}$ and $\boldsymbol{\theta}$ contains the weights of neural network. Besides, the replay buffer generated by the repeated interactions can be utilized to update $\boldsymbol{\theta}$.

To improve the learning accuracy and efficiency of Q-function, two neural networks with the same three fully connected (FC) layers are adopted, namely, evaluation NN and target NN.[1] During the training process, a minibatch $\mathcal{B} = \{\mathbf{e}_i\}_{1 \le i \le Y}$ is sampled from the replay buffer. For each experience $\mathbf{e}_i = \{\mathbf{s}_i; \mathbf{a}_i; r_i; \mathbf{s}_{i+1}\}$, the evaluation network takes state $\mathbf{s}_i$ as the input and outputs the evaluated Q-function values, while the target network takes the $\mathbf{s}_{i+1}$ as the input and generates the target Q-function values for the i-th experience. The weight of evaluation network is updated to minimize the loss function, which is given by

$$Loss(\boldsymbol{\theta}^E) = \mathbb{E}_{\mathbf{e}_i \in \mathcal{B}} \left[\left(q_i - Q(\mathbf{s}_i, \mathbf{a}_i; \boldsymbol{\theta}^E) \right)^2 \right], \tag{9}$$

where $q_i = r_i + \eta \max_{\mathbf{a}'} Q(\mathbf{s}_{i+1}, \mathbf{a}'; \boldsymbol{\theta}^T)$ represents target Q-function values, η is a discount factor, $\boldsymbol{\theta}^E$ and $\boldsymbol{\theta}^T$ denote the weights of evaluation and target NN, respectively. The weights $\boldsymbol{\theta}^E$ are optimized by Adam method [14]. After fixed training steps L, we set $\boldsymbol{\theta}^T = \boldsymbol{\theta}^E$ to stabilize the learning process.

In particular, to simplify the selection of P_A, a solver NN is firstly proposed to replace traditional bisection method in solving $P_{so} = \varepsilon$, which takes P_j and R_s as input and outputs the estimated P_A.[2] Moreover, transfer learning is

[1] To stabilize learning process at the beginning, target NN should have the same structure with evaluation NN, which is utilized to duplicate evaluation NN every fixed time and generate steady target Q-function values [13].

[2] The time complexity of our considered solver NN is $\mathcal{O}(2n_1 + n_1 n_2 + n_2)$, whereby n_i denotes the nodes number of i-th layer [15], while the complexity of bisection method is $\mathcal{O}\left(\log_2(N_{P_A})\right)$, where N_{P_A} is the size of available transmit power set.

introduced to initialize the weights $\boldsymbol{\theta}^S$ of solver NN, which can simplify the network training and accelerate the algorithm convergence [7]. The details of our proposed DeepAntiJam are shown in **Algorithm 1**.[3]

Algorithm 1 DeepAntiJam

```
Initialize networks weights $\boldsymbol{\theta}^E$, $\boldsymbol{\theta}^T$, and $\boldsymbol{\theta}^S$ with transfer learning.
Set $\varepsilon$, $\eta$, buffer size $Z$, $Y$, $\mathcal{B} = \emptyset$, steps $L$, and $\triangle P$.
Evaluate jamming power $P_J^0$, $\mathbf{s}^1 = \{P_J^0\}$.
for $k = 1, 2, \cdots$ do
    Set $\mathbf{s}^k$ as input of the evaluation NN and obtain the output $Q(\mathbf{s}^k, \mathbf{a}; \boldsymbol{\theta}^E)$.
    Select $R_s^k$ via greedy policy for Alice.
    Set $\{R_s^k, P_J^{k-1}\}$ as the input to solver NN, obtain $P_A^k$.
    Calculate the secrecy outage probability $P_{so}$.
    while $P_{so} > \varepsilon$ or $P_A^k > P_A^{max}$ do
        $P_A^k = P_A^k - \triangle P$, and calculate $P_{so}$.
    end while
    Calculate the reward $r^k$ based on (1).
    Evaluate jamming power $P_J^k$, $\mathbf{s}^{k+1} = \{P_J^k\}$.
    Formulate an experience: $\mathbf{e}^k = \{\mathbf{s}^k, \mathbf{a}^k, r^k; \mathbf{s}^{k+1}\}$.
    $\mathcal{B} \leftarrow \mathcal{B} \cup \mathbf{e}^k$, $\mathcal{B} = \{\mathbf{e}^{k-Z}, \mathbf{e}^{k-Z+1}, \cdots, \mathbf{e}^k\}$.
    if $k > Y$ then
        Randomly sample $Y$ experiences from reply buffer.
        Update $q_i = r_i + \eta \max_{\mathbf{a}'} Q(\mathbf{s}_{i+1}, \mathbf{a}'; \boldsymbol{\theta}^T)$,$1 \leq i \leq Y$.
        Optimize $\boldsymbol{\theta}^E$ with Adam by minimizing (9).
        if $k \equiv 0 \pmod{L}$ then
            Set $\boldsymbol{\theta}^T = \boldsymbol{\theta}^E$.
        end if
    end if
end for
```

IV. Simulation Results

In this section, simulation results are presented to evaluate the effectiveness of our proposed scheme. In particular, there are $M = 3$ eavesdroppers, and similar to [3] and [10], we set $R_t = 4$ bps/Hz, $\eta = 0.9$, $\sigma_{AB}^2 = 0.1$, $\sigma_{JB}^2 = 0.01$, $\sigma_{AE_1}^2 = 0.0001$, $\sigma_{AE_2}^2 = 0.0002$, $\sigma_{AE_3}^2 = 0.001$, $\sigma_{JE_1}^2 = 0.05$, $\sigma_{JE_2}^2 = 0.04$, and $\sigma_{JE_3}^2 = 0.003$. The maximal power of transmitter P_A^{max} is 30W, and the sets $\mathcal{P}_J$ and $\mathcal{R}_s$ are set to be {0, 0.5, 1, . . . , 6} and {0.8, 1, 1.2, . . . , 3.8}, respectively. Without loss of generality, the noise powers at each receiver are assumed to be same and equal to 0.01, i.e., $\sigma_B^2 = \sigma_{E_m}^2 = 0.01$.

In terms of simulation environment, all simulations are carried out on a desktop with 12-core 3.6Ghz Intel i7-12700K microprocessor. Specifically, the proposed DeepAntiJam architecture is implemented by using the Torch library on the Pytorch platform (version 1.11.0) [16] in Python 3.9 version [17]. In terms of simulation parameters, the hyperparameters for DeepAntiJam are as follows, the number of neurons in the hidden layers of two considered evaluation NNs are (64) and (64, 128, 64), respectively, and the number of neurons in the hidden layers of solver NN is (64, 16). The batch size is 64, and the learning rate of all NN is 0.001. The sigmoid function is used as the activation function.

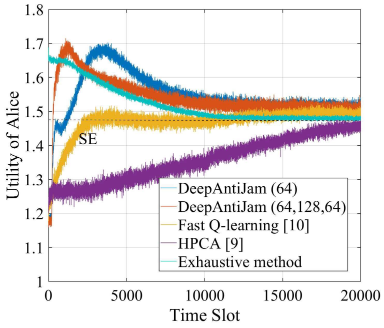

Fig. 3. Performance comparison of HPCA, fast Q-learning, exhaustive method, and our proposed DeepAntiJam against Q-learning based jammer.

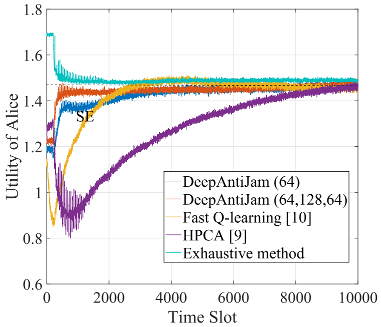

Fig. 4. Performance comparison of HPCA, fast Q-learning, exhaustive method, and our proposed DeepAntiJam against deep RL-based jammer.

To show the superiority of our proposed DeepAntiJam, Fig. 3 is presented to illustrate Alice's utility of four schemes, namely, HPCA [9], fast Q-learning [10], exhaustive method, and our proposed DeepAntiJam. It is apparent that our proposed DeepAntiJam achieves a better jamming resistance performance with larger cumulative utility than HPCA and fast Q-learning. Though all the three learning schemes converge to SE at last, due to its outstanding learning ability compared to HPCA and fast Q-learning, our proposed scheme finds the suitable transmission strategy quickly. As a result, DeepAntiJam achieves a higher utility at the beginning and its utility degrades slowly, which results in a higher cumulative utility compared to HPCA and fast Q-learning. In addition, although the exhaustive method can achieve the similar performance as DeepAntiJam, it is impractical to resist dynamic jamming due to its high implementation complexity.[4]

Figure 4 depicts the utility of Alice against deep RL-based jammer which utilizes deep RL with network (64, 128, 64) to obtain its jamming strategy. Compared to Fig. 3, due to the fact that deep RL-based smart jammer can learn a better jamming strategy than Q-learning based smart jammer, the utilities achieved by the four anti-jamming schemes degrades obviously. However, SE can always be realized by all the four schemes. In addition, our proposed DeepAntiJam can obtain a significant enhancement of secrecy throughput at the beginning

[3] The complexity of our proposed DeepAntiJam is mainly determined by the evaluation NN and solver NN, which is $\mathcal{O}(m_1 + m_1 m_2 + m_2 m_3 + m_3|\mathcal{R}_s| + 2n_1 + n_1 n_2 + n_2)$ with m_i denoting the number of nodes in the i-th layer of evaluation NN and $|\mathcal{R}_s|$ denoting the size of secrecy rate set.

[4] The computational complexity of exhaustive method is $\mathcal{O}\big(|\mathcal{R}_s| \log_2(N_{P_A})\big)$, where N_{P_A} denotes the size of available transmit power set.

compared to HPCA and fast Q-learning, which guarantees a larger cumulative utility. Moreover, it is apparent that the convergence speed and anti-jamming ability of our proposed DeepAntiJam are determined by the dimension of evaluation NN, where more complicated evaluation NN will lead to faster convergence and stronger anti-jamming ability.

V. Conclusion

In this letter, a deep RL-assisted anti-jamming transmission scheme, called DeepAntiJam, was proposed to update the transmission strategy of legitimate transceiver under the coexistence of one smart jammer and multiple passive eavesdroppers. Firstly, the secure transmission was formulated as a Stackelberg game, and the existence and uniqueness of its equilibrium were proved. To improve the anti-jamming ability and arrive at SE, DeepAntiJam was developed to overcome the threat of learning-assisted jammer. Moreover, transfer learning was introduced into initialization to accelerate the convergence performance. Simulation results validate that our proposed secure transmission methodology can improve the secrecy throughput significantly in the presence of learning-based jammer.

Appendix
Proof of Lemma 1

The first-order derivative of P_{co} with respect to P_A can be obtained, given by

$$\begin{aligned}\frac{\partial P_{co}}{\partial P_A} &= \frac{\partial\left[1 - f_1(P_A) f_2(P_A)\right]}{\partial P_A}, \\ &= -f_2(P_A)\frac{\partial f_1(P_A)}{\partial P_A} - f_1(P_A)\frac{\partial f_2(P_A)}{\partial P_A}, \qquad (10)\end{aligned}$$

where $f_1(P_A)$ and $f_2(P_A)$ are denoted as $f_1(P_A) = \exp\left(-\frac{\gamma_1 \sigma_B^2}{P_A \sigma_{AB}^2}\right)$ and $f_2(P_A) = \frac{P_A \sigma_{AB}^2}{P_A \sigma_{AB}^2 + \gamma_1 P_J \sigma_{JB}^2}$, respectively. Since $\frac{\partial f_1(P_A)}{\partial P_A} > 0$ and $\frac{\partial f_2(P_A)}{\partial P_A} > 0$, we can obtain that $\frac{\partial P_{co}}{\partial P_A} < 0$.

The first-order derivative of P_{so} with respect to P_A can be expressed as

$$\begin{aligned}\frac{\partial P_{so}}{\partial P_A} &= \frac{\partial\left[1 - \prod_{m=1}^{M} g_m\right]}{\partial P_A}, \\ &= -\sum_{k=1}^{M}\left(\frac{\partial g_k}{\partial P_A} \prod_{m=1,\ldots,k-1,k+1,\ldots,M} g_m\right), \qquad (11)\end{aligned}$$

where $g_m = \mathbf{P}\left\{\gamma_{E_m} \le \gamma_2\right\} = 1 - \exp\left(-\frac{\gamma_2 \sigma_{E_m}^2}{P_A \sigma_{AE_m}^2}\right) \frac{P_A \sigma_{AE_m}^2}{P_A \sigma_{AE_m}^2 + \gamma_2 P_J \sigma_{JE_m}^2} > 0$. Accordingly, we have $\frac{\partial g_k}{\partial P_A} < 0$, thus $\frac{\partial P_{so}}{\partial P_A} > 0$.

The first-order derivative of P_{so} with respect to P_A can be derived as

$$\begin{aligned}\frac{\partial P_{so}}{\partial R_s} &= \frac{\partial\left[1 - \prod_{m=1}^{M} g_m\right]}{\partial R_s}, \\ &= -\sum_{k=1}^{M}\left(\frac{\partial g_k}{\partial R_s} \prod_{m=1,\ldots,k-1,k+1,\ldots,M} g_m\right). \qquad (12)\end{aligned}$$

Then, we take first-order derivative of g_m with respect to R_s, which can be derived as

$$\begin{aligned}\frac{\partial g_m}{\partial R_s} &= \frac{\partial\left[1 - \exp\left(-\frac{\gamma_2 \sigma_{E_m}^2}{P_A \sigma_{AE_m}^2}\right) \frac{P_A \sigma_{AE_m}^2}{P_A \sigma_{AE_m}^2 + \gamma_2 P_J \sigma_{JE_m}^2}\right]}{\partial R_s}, \\ &= -\frac{\partial \exp\left(-\frac{\gamma_2 \sigma_{E_m}^2}{P_A \sigma_{AE_m}^2}\right)}{\partial R_s} \frac{P_A \sigma_{AE_m}^2}{P_A \sigma_{AE_m}^2 + \gamma_2 P_J \sigma_{JE_m}^2} \\ &\quad - \frac{\partial \frac{P_A \sigma_{AE_m}^2}{P_A \sigma_{AE_m}^2 + \gamma_2 P_J \sigma_{JE_m}^2}}{\partial R_s} \exp\left(-\frac{\gamma_2 \sigma_{E_m}^2}{P_A \sigma_{AE_m}^2}\right), \\ &= -\frac{\exp\left(-\frac{\gamma_2 \sigma_{E_m}^2}{P_A \sigma_{AE_m}^2}\right) \sigma_{E_m}^2 2^{R_t - R_s} \ln 2}{P_A \sigma_{AE_m}^2} \\ &\quad - \frac{P_A \sigma_{AE_m}^2 P_J \sigma_{JE_m}^2 2^{R_t - R_s} \ln 2}{\left(P_A \sigma_{AE_m}^2 + \gamma_2 P_J \sigma_{JE_m}^2\right)^2} < 0. \qquad (13)\end{aligned}$$

Substituting (13) into (12), we have $\frac{\partial P_{so}}{\partial R_s} > 0$.

References

[1] J. M. Hamamreh, H. M. Furqan, and H. Arslan, "Classifications and applications of physical layer security techniques for confidentiality: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 2, pp. 1773–1828, 2nd Quart., 2019.

[2] H. Fang, L. Xu, Y. Zou, X. Wang, and K.-K. R. Choo, "Three-stage Stackelberg game for defending against full-duplex active eavesdropping attacks in cooperative communication," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 10788–10799, Nov. 2018.

[3] W. Wang, K. C. Teh, K. H. Li, and S. Luo, "On the impact of adaptive eavesdroppers in multi-antenna cellular networks," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 2, pp. 269–279, Feb. 2018.

[4] L. Jia *et al.*, "Stackelberg game approaches for anti-jamming defence in wireless networks," *IEEE Wireless Commun.*, vol. 25, no. 6, pp. 120–128, Dec. 2018.

[5] 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.

[6] L. Xiao, T. Chen, J. Liu, and H. Dai, "Anti-jamming transmission Stackelberg game with observation errors," *IEEE Commun. Lett.*, vol. 19, no. 6, pp. 949–952, Jun. 2015.

[7] 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.

[8] D. He, H. Wang, and H. Zhou, "Learning-based secure communication against active eavesdropper in dynamic environment," *IET Commun.*, vol. 13, no. 15, pp. 2235–2242, Sep. 2019.

[9] L. Jia, F. Yao, Y. Sun, Y. Xu, S. Feng, and A. Anpalagan, "A hierarchical learning solution for anti-jamming Stackelberg game with discrete power strategies," *IEEE Wireless Commun. Lett.*, vol. 6, no. 6, pp. 818–821, Dec. 2017.

[10] L. Xiao, Y. Li, C. Dai, H. Dai, and H. V. Poor, "Reinforcement learning-based NOMA power allocation in the presence of smart jamming," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3377–3389, Apr. 2018.

[11] C. Wang and H.-M. Wang, "On the secrecy throughput maximization for MISO cognitive radio network in slow fading channels," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 11, pp. 1814–1827, Nov. 2014.

[12] R. Lucchetti, F. Mignanego, and G. Pieri, "Existence theorems of equilibrium points in Stackelberg," *Optimization*, vol. 18, no. 6, pp. 857–866, 1987.

[13] Z. Ji, A. K. Kiani, Z. Qin, and R. Ahmad, "Power optimization in device-to-device communications: A deep reinforcement learning approach with dynamic reward," *IEEE Wireless Commun. Lett.*, vol. 10, no. 3, pp. 508–511, Mar. 2021.

[14] D. P. kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, San Diego, CA, USA, May 2015, pp. 1–15.

[15] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5353–5360.

[16] *Pytorch*. Accessed: Apr. 25, 2022. [Online]. Available: https://pytorch.org/

[17] *Python*. Accessed: Apr. 25, 2022. [Online]. Available: https://www.python.org/