

A Proactive Eavesdropping Game in MIMO Systems Based on Multiagent Deep Reinforcement Learning

Delin Guo^{ID}, Hui Ding, Lan Tang^{ID}, *Member, IEEE*, Xinggan Zhang^{ID}, Lvxi Yang^{ID}, *Senior Member, IEEE*, and Ying-Chang Liang^{ID}, *Fellow, IEEE*

Abstract—This paper considers an adversarial scenario between a legitimate eavesdropper and a suspicious communication pair. All three nodes are equipped with multiple antennas. The eavesdropper, which operates in a full-duplex model, aims to wiretap the dubious communication pair via proactive jamming. On the other hand, the suspicious transmitter, which can send artificial noise (AN) to disturb the wiretap channel, aims to guarantee secrecy. More specifically, the eavesdropper adjusts jamming power to enhance the wiretap rate, while the suspicious transmitter jointly adapts the transmit power and noise power against the eavesdropping. Considering the partial observation and complicated interactions between the eavesdropper and the suspicious pair in unknown system dynamics, we model the problem as an imperfect-information stochastic game. To approach the Nash equilibrium solution of the eavesdropping game, we develop a multi-agent reinforcement learning (MARL) algorithm, termed neural fictitious self-play with soft actor-critic (NFSP-SAC), by combining the fictitious self-play (FSP) with a deep reinforcement learning algorithm, SAC. The introduction of SAC enables FSP to handle the problems with continuous and high dimension observation and action space. The simulation results demonstrate that the power allocation policies learned by our method empirically converge to a Nash equilibrium, while the compared reinforcement learning algorithms suffer from severe fluctuations during the learning process.

Index Terms—Physical layer security, proactive eavesdropping, stochastic game, multiagent reinforcement learning.

Manuscript received 29 June 2021; revised 6 February 2022 and 4 April 2022; accepted 18 April 2022. Date of publication 6 May 2022; date of current version 11 November 2022. This work was supported in part by the National Natural Science Foundation of China under Grant U1936201, Grant 62072229, and Grant 62071220 and in part by the State Key Laboratory of Air Traffic Management System and Technology under Grant SKLATM202109. The associate editor coordinating the review of this article and approving it for publication was K. Kansanen. (*Corresponding author: Lan Tang.*)

Delin Guo, Lan Tang, and Xinggan Zhang are with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210023, China, and also with the State Key Laboratory of Air Traffic Management Systems and Technology, Nanjing 210008, China (e-mail: tanglan@nju.edu.cn).

Hui Ding is with the State Key Laboratory of Air Traffic Management Systems and Technology, Nanjing 210093, China.

Lvxi Yang is with the School of Information Science and Engineering, Southeast University, Nanjing 211189, China.

Ying-Chang Liang is with the Center for Intelligent Networking and Communications (CINC), University of Electronic Science and Technology of China (UESTC), Chengdu 611731, China, and also with the Yangtze Delta Region Institute, University of Electronic Science and Technology of China, Huzhou 313001, China (e-mail: liangyc@ieee.org).

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TWC.2022.3170308>.

Digital Object Identifier 10.1109/TWC.2022.3170308

I. INTRODUCTION

A. Backgrounds and Motivations

IN RECENT decades, wireless communication has played an essential role in many aspects of social life, by providing an efficient and convenient way for connecting people. However, due to the broadcast nature of the wireless channel, the information security in wireless communications is susceptible to potential security vulnerabilities. Hence, how establishing reliable and secure connections in wireless communication systems is a challenging task. Conventionally, the data transmission security is ensured by applying cryptographic methods at the higher layer of communication protocols [1]. Due to the significant improvement of wireless computing power, coupled with the challenges in distribution and management of wireless keys, physical layer security has attracted lots of interest [2]–[5]. Typical technologies of physical layer security were concluded in [6], such as secure resource allocation, secure beamforming and precoding, antenna/node selection and cooperation, and joint strategies of several above approaches.

The works mentioned above focus on guaranteeing the security of wireless communications by treating the communication users to be legal while assuming the eavesdropper as vicious invaders. However, wireless communication links could also be utilized by criminals to facilitate their illegal activities. Especially with the development of infrastructure-free communications, the illegal information may not be transmitted through core infrastructure, such that it is hard to be monitored by conventional surveillance methods [7]. Therefore, to improve public security, authorized parties (e.g., government agencies) have strong need to monitor the suspicious communication links.

Passive eavesdropping is a straightforward way in wireless wiretaps, which simply listens to the dubious wireless connection via a silent eavesdropper. However, with the application of high-frequency bands and massive multiple-input multiple-output (MIMO) in the 5th generation (5G) networks, the transmit beam is narrowed. When legal eavesdroppers are deployed far away from the suspicious transmitter, the passive eavesdropping is inefficient. To improve the monitoring efficiency, proactive eavesdropping via jamming was firstly proposed in [7] and [8], where the full-duplex eavesdropper reduced the transmission rate of dubious links by transmitting jamming signals to enhance the wiretap efficiency.

Subsequently, the jamming-assisted proactive eavesdropping approaches were studied extensively in [9]–[13]. Another novel approach, proactive jamming via spoofing relay was proposed in [14], where three spoofing relay policies were investigated to achieve the maximum wiretap rate. Reference [15] proposed a relay-assisted eavesdropping scheme, which maximized the eavesdropping rate by coordinating the central monitor, spoofing relays, and the jammer. Moreover, [16] and [17] proposed the pilot-contamination-based methods in time-division duplex (TDD) systems, which made the suspicious transmitter misestimate the channel by utilizing pilot contamination during the reverse training phase.

Considering the potential countermeasures taken by opponents¹ in the physical layer security problems, some works based on the game theory were proposed. References [18] and [19] studied the game between a half-duplex eavesdropper (which acted as either a passive eavesdropper or an active jammer) and a communication pair in the MIMO system, in which the Nash equilibrium solution was considered. Reference [20] studied the anti-jamming problem where the attack type of the jammer is unknown. Therefore, the problem is modeled as a Bayesian Game where the utility of communication nodes are computed based on their individual beliefs. Reference [21] investigated the defending against full-duplex proactive eavesdropping in the relay-assisted communication system by modeling the countermeasure problem as a three-stage Stackelberg game, where the game operates among a full-duplex eavesdropper, all relays, and a wireless link. To combat the full-duplex eavesdropper in the multi-channel wireless system, [22] formulated the problem as a two-stage hierarchical game.

In recent years, considering unknown system dynamics, some works modeled the security problems as stochastic games [23], which were solved by reinforcement learning (RL) [24] methods. Reference [25] formulated the problem of resisting jamming and interference as a non-zero-sum game, which was solved by the fast deep-Q-network (DQN) [26]. Reference [27] modeled the defense and attack in the unmanned aerial vehicle (UAV) communication system as a prospect-theory-based game and solved it by applying Q-learning [24], DQN [28], and Win-or-Learn Fast Policy Hill Climbing (WoLF-PHC) [29]. Reference [30] proposed two multi-agent reinforcement learning (MARL) [23] algorithms, termed minimax post-decision state (minimax-PDS) and WoLF-PDS, to obtain the Nash equilibrium solution of security games with asymmetry information.

The methods mentioned above and some other existing algorithms for solving stochastic games have some non-negligible weaknesses. 1) Single-agent RL algorithms (e.g., Q-learning, DQN, and fast DQN) are rational but not convergent since they violate the stationary assumption required for the convergence when applying these methods in multi-agent tasks, where “rational” and “convergent” are two desirable properties for multi-agent learning algorithms introduced by [29]: A learning algorithm is called rational if given the stable policies of other agents, the agent’s policy using this algorithm converges to

the best response (i.e., the optimal counter-policy to opponents’ behavior) to other agents, and called convergent if this algorithm will necessarily converge to a stationary state. These two properties are described in the section III in more detail. 2) minimax-Q [31], minimax-PDS, and Nash-Q [32] are convergent but not rational, and minimax-based methods are only suitable for zero-sum stochastic games. 3) WoLF-PHC and WoLF-PDS are rational, and it is empirically convergent in some simple games (e.g., matrix games), but they are difficult to converge in more complex games. Moreover, these table-based methods can not cope with the continuous or high-dimensionality problems, which are typically encountered in practical communication systems. 4) Reference [33] proposed the NFSP-DQN by combining fictitious self-play (FSP) [34] with DQN, and it approached the Nash equilibriums in Leduc poker and Limit Texas Hold’em, two poker games with imperfect information. However, NFSP-DQN is only suitable for problems with low dimensional and discrete action space. Besides, because DQN can only choose action deterministically, NFSP-DQN can not solve games with Mixed strategy Nash equilibrium (MSNE), even very simple games such as Rock-Paper-Scissor.

B. Main Contributions

In this paper, we aim to improve the eavesdropping rate of the legitimate eavesdropper against a malicious communication pair who tries to protect suspicious information’s secrecy. We model the problem as a general-sum imperfect-information stochastic game. To approach the Nash equilibrium of this stochastic game and overcome the disadvantages of the existing methods mentioned above, we enhance the FSP method by combining it with the soft actor-critic (SAC) algorithm [35]. The main contributions of this paper are as follows:

- Firstly, we consider the adversarial problem between a legitimate eavesdropper and a suspicious communication pair in the MIMO system. The full-duplex eavesdropper aims to enhance the wiretap rate via proactive jamming, while the suspicious transmitter combats the eavesdropping by sending artificial noise (AN) signal and adjusting the transmit power of information. Moreover, we assume imperfect self-interference cancellation at the eavesdropper. Considering the partial observations and the interrelationship between both sides, we model the problem as an imperfect-information stochastic game, and then we obtain the Nash equilibrium policy profile.
- Secondly, to approach the Nash equilibrium of the modeled stochastic game, we develop a novel algorithm, termed neural fictitious self-play with soft actor-critic (NFSP-SAC), by introducing the SAC into FSP. In particular, the NFSP-SAC agent maintains two policies, the best response and the averaged policy, both of which are parameterized by deep neural networks (DNN). The agent chooses actions according to the mixture policy of the best response and the averaged policy. The best response is trained by SAC from all historical data stored in a memory buffer, which is collected by interacting with the agent’s opponents. The averaged policy is learned

¹If we stand for the communication link, the opponent is the eavesdropper; otherwise, the opponent is the communication link.

from the historical behaviors only executed by the best response policy via supervised learning.

- Finally, the performance of NFSP-SAC is verified by extensive simulation results. It is shown that our algorithm can asymptotically converge to an equilibrium state, while the benchmark methods, such as the normal SAC and the multi-agent RL method WoLF-PPO [36], suffer from severe fluctuations and performance degradations during the learning process.

C. Related Works

To protect transmissions between the transmitter and intended receiver from potential eavesdroppers, artificial noise (AN) is generated at the transmitter and is injected into the null-subspace of the main channel of the communication user pair [2], [37], [38]. AN is utilized to impair the eavesdropper's channel, while not affecting the channel between the communication user pair. On the other hand, proactive eavesdropping via jamming was firstly proposed by [7] and [8] to realize the monitoring of illegal users. Some works considered both eavesdropping and anti-eavesdropping, among which [18] and [39] studied the system model similar to ours. In [18], a jamming game in a MIMO system was studied, while the eavesdropper operates in half-duplex mode, which was different from ours. In [39], an eavesdropping game in a single-input single-output system was investigated, where the main difference was that Alice cannot send AN. Besides, both of [18] and [39] did not solve problems in the framework of learning. Some work (e.g., [27], [30]) modeled the security problems in wireless communications as stochastic games, which were solved by RL-based methods, but the considered problems were widely different from ours.

In this paper, we combine FSP with SAC to solve the considered eavesdropping game. Fictitious play (FP) [40] is a standard tool in game theory, which maintains two policies: the best response and an average policy. However, this method is only suitable for small-scale games. Reference [20] applied FP in the anti-jamming game where the type of the jammer is unknown. Reference [41] enhanced FP by introducing deep neural networks, and although there was no convergence guarantee, it still achieved the Nash equilibrium in a game with fifty agents. Reference [42] proved the convergence of FP under appropriate assumptions. Reference [34] proposed a machine learning framework named FSP based on the generalized weakened FP [43], where the best response was learned by RL algorithms, and the averaged policy was learned by supervised learning. Subsequently, [33] developed NFSP-DQN by combining FSP with DQN to solve stochastic games with imperfect information. Reference [44] showed that NFSP-DQN achieved the safest performance in an autonomous vehicle control game with a non-zero-sum setting. Reference [45] applied FSP to solve real-time strategy games, where the policy empirically converged to the Nash equilibrium. In this paper, SAC is applied to FSP to learn the best response. We derive the policy gradient of the objective in SAC with a parameterized policy. Updating the policy by gradient ascent using the derived policy gradient, the policy provably converges to at least a local optimum with some strict conditions according to [46], [47].

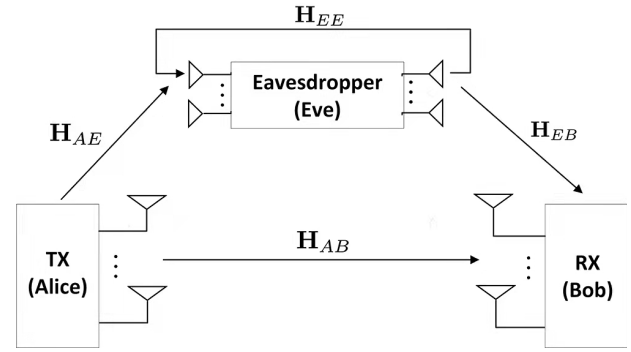


Fig. 1. System model of a three-node MIMO wiretap system.

The rest of this paper is organized as follows. Some necessary backgrounds and definitions are introduced in Section II. The system model is presented in Section III. In Section IV, the eavesdropping problem is formulated as an imperfect-information stochastic game. The proposed algorithm, NFSP-SAC, which is used for tackling this game, is elaborated in Section VI. The numerical simulations are provided in Section VI. Finally, Section VII concludes this work.

Notation: We use bold upper case letters to denote matrices, bold lower case letters to denote vectors and lower case letters to denote scalars. $(\cdot)^H$ and $(\cdot)^{-1}$ denote conjugate transpose operator and matrix inverse, respectively. $\det(\cdot)$ denotes the determinant of matrix. $\mathbf{I}_k$ is the identity matrix of size k .

II. SYSTEM MODEL

As shown in Fig. 1, we investigate a three-node MIMO wiretap system, where a full-duplex legitimate eavesdropper (Eve) aims to monitor the suspicious wireless link between a transmitter (Alice) and a receiver (Bob), by proactive jamming. The suspicious nodes Alice and Bob are equipped with N_A and N_B antennas, respectively. Similar to [3], [37], and [38], we assume that Alice can prevent Eve from eavesdropping by transmitting AN signals. For the legitimate monitor, in order to enable simultaneous eavesdropping and jamming, it is equipped with two sets of antennas, one set with N_E^r antennas for receiving (eavesdropping) and the other set with N_E^t antennas for transmitting (jamming). Due to co-located antennas, severe self-interference (SI) may occur between Eve's transmit and receive antennas. Although some techniques (e.g., [48], [49]) have been developed to mitigate the SI, the residual SI still exists due to some practical constraints [3], [13]. Similar to [3] and [48], we model the residual SI channel at Eve as $\sqrt{\rho}\mathbf{H}_{EE}$, where $\mathbf{H}_{EE} \in \mathbb{C}^{N_E^r \times N_E^t}$ denotes the imaginary fading loop channel and ρ parameterizes the effect of passive SI suppression. We assume that N_A is larger than N_B , so that Alice has additional spatial degrees of freedom for transmitting AN signals and doesn't need to sacrifice the spatial freedom of sending data information. This assumption could be considered in scenarios where Alice is a base station or some fixed access point with large-scale antennas, while Bob is a user terminals with less antennas. We do not assume the relationship between N_E^r and (N_B, N_A) . If the hardware complexity allows, increasing N_E^r would probably help to improve the monitoring performance. Let $\mathbf{H}_{AB} \in \mathbb{C}^{N_B \times N_A}$, $\mathbf{H}_{AE} \in \mathbb{C}^{N_E^r \times N_A}$ and $\mathbf{H}_{EB} \in \mathbb{C}^{N_B \times N_E^t}$ denote the

Alice-to-Bob, Alice-to-Eve and Eve-to-Bob channels, respectively. We suppose that all elements in the three channel matrices obey the complex Gaussian distribution $\mathcal{CN}(0,1)$, and channels suffer from Markov block fading. That is, the channel coefficients keep constant over each transmission time slot (TS), but may vary according to Markov model between different TS.

A. Signal Model

The signal transmitted by Alice consists of an information signal and an AN signal, where the AN is used to prevent being eavesdropped on. The information and AN signal transmitted by Alice are denoted by $\mathbf{x}_A^{IS} \in \mathbb{C}^{N_B \times 1}$ and $\mathbf{x}_A^{AN} \in \mathbb{C}^{(N_A - N_B) \times 1}$, respectively. To perform this transmission, an $N_A \times N_A$ precoding matrix is designed as

$$\mathbf{W}_A = [\mathbf{W}_A^{IS} \ \mathbf{W}_A^{AN}], \quad (1)$$

where $\mathbf{W}_A^{IS} \in \mathbb{C}^{N_A \times N_B}$ is used to precode $\mathbf{x}_A^{IS}$, and $\mathbf{W}_A^{AN} \in \mathbb{C}^{N_A \times (N_A - N_B)}$ is used to precode $\mathbf{x}_A^{AN}$. In order to ensure that the AN signal does not interfere with the information signal, we take $\mathbf{W}_A^{IS}$ and $\mathbf{W}_A^{AN}$ as disjoint sets of the right singular vectors of $\mathbf{H}_{AB}$ [8], [18]. More clearly, $\mathbf{W}_A^{IS}$ is comprised of the right singular vectors corresponding to the N_B non-zero singular values, while $\mathbf{W}_A^{AN}$ is comprised of the remaining $N_A - N_B$ right singular vectors. As such, $\mathbf{W}_A^{AN}$ lies in the null-space of $\mathbf{H}_{AB}$, i.e., $\mathbf{H}_{AB}\mathbf{W}_A^{AN} = 0$. The total signal transmitted by Alice, $\mathbf{x}_A \in \mathbb{C}^{N_A \times 1}$, is given by [38]

$$\begin{aligned} \mathbf{x}_A &= [\mathbf{W}_A^{IS} \ \mathbf{W}_A^{AN}] \begin{bmatrix} \mathbf{x}_A^{IS} \\ \mathbf{x}_A^{AN} \end{bmatrix} \\ &= \mathbf{W}_A^{IS} \mathbf{x}_A^{IS} + \mathbf{W}_A^{AN} \mathbf{x}_A^{AN}. \end{aligned} \quad (2)$$

Then $\mathbf{x}_A$ is transmitted to Bob via the channel $\mathbf{H}_{AB}$. Besides, the monitor Eve is transmitting a jamming signal $\mathbf{x}_E \in \mathbb{C}^{N_E \times 1}$ to Bob via $\mathbf{H}_{EB}$ to reduce the received signal noise ratio (SNR) of Bob. Because $\mathbf{H}_{AB}\mathbf{W}_A^{AN} = 0$, the received signal at Bob is given by

$$\mathbf{y}_B = \sqrt{g_{AB}}\mathbf{H}_{AB}\mathbf{W}_A^{IS}\mathbf{x}_A^{IS} + \sqrt{g_{EB}}\mathbf{H}_{EB}\mathbf{x}_E + \mathbf{n}_B, \quad (3)$$

where $g_{AB} = d_{AB}^{-\lambda}$ (d_{AB} is the distance between Alice and Bob) and $g_{EB} = d_{EB}^{-\lambda}$ (d_{EB} is the distance between Eve and Bob) are the path losses with exponent λ , respectively, and $\mathbf{n}_B$ is the complex Gaussian noise vector with zero-mean and covariance $\mathbb{E}\{\mathbf{n}_B\mathbf{n}_B^H\} = \sigma_B^2\mathbf{I}_{N_B}$.

On the other hand, due to the broadcast nature of wireless channel, the signal transmitted by Alice, $\mathbf{x}_A$, can also be received by Eve via $\mathbf{H}_{AE}$. The received signal at the Eve is given by

$$\begin{aligned} \mathbf{y}_E &= \sqrt{g_{AE}}\mathbf{H}_{AE}\mathbf{W}_A^{IS}\mathbf{x}_A^{IS} + \sqrt{g_{AE}}\mathbf{H}_{AE}\mathbf{W}_A^{AN}\mathbf{x}_A^{AN} \\ &\quad + \sqrt{\rho}\mathbf{H}_{EE}\mathbf{x}_E + \mathbf{n}_E, \end{aligned} \quad (4)$$

where $g_{AE} = d_{AE}^{-\lambda}$ (d_{AE} is the distance between Alice and Eve), $\mathbf{n}_E$ is the complex Gaussian noise with $\mathbb{E}\{\mathbf{n}_E\mathbf{n}_E^H\} = \sigma_E^2\mathbf{I}_{N_E}$.

The covariance matrices of the received information signals at Bob and Eve are

$$\mathbf{K}_B = g_{AB}\mathbf{H}_{AB}\mathbf{W}_A^{IS}\mathbf{Q}_{\mathbf{x}_A^{IS}}(\mathbf{W}_A^{IS})^H\mathbf{H}_{AB}^H, \quad (5)$$

and

$$\mathbf{S}_E = g_{AE}\mathbf{H}_{AE}\mathbf{W}_A^{IS}\mathbf{Q}_{\mathbf{x}_A^{IS}}(\mathbf{W}_A^{IS})^H\mathbf{H}_{AE}^H, \quad (6)$$

respectively, where $\mathbf{Q}_{\mathbf{x}_A^{IS}} = \mathbb{E}\{\mathbf{x}_A^{IS}(\mathbf{x}_A^{IS})^H\}$. Besides, the covariance matrices of the received interference-plus-noise (IN) at Bob and Eve are given by

$$\mathbf{K}_B = g_{EB}\mathbf{H}_{EB}\mathbf{Q}_{\mathbf{x}_E}\mathbf{H}_{EB}^H + \sigma_B^2\mathbf{I}_{N_B}, \quad (7)$$

and

$$\begin{aligned} \mathbf{K}_E &= g_{AE}\mathbf{H}_{AE}\mathbf{W}_A^{AN}\mathbf{Q}_{\mathbf{x}_A^{AN}}(\mathbf{W}_A^{AN})^H\mathbf{H}_{AE}^H \\ &\quad + \rho\mathbf{H}_{EE}\mathbf{Q}_{\mathbf{x}_E}\mathbf{H}_{EE}^H + \sigma_E^2\mathbf{I}_{N_E}, \end{aligned} \quad (8)$$

respectively, where $\mathbf{Q}_{\mathbf{x}_E} = \mathbb{E}\{\mathbf{x}_E\mathbf{x}_E^H\}$ and $\mathbf{Q}_{\mathbf{x}_A^{AN}} = \mathbb{E}\{\mathbf{x}_A^{AN}(\mathbf{x}_A^{AN})^H\}$.

The total transmit power of Alice is denoted by P_A^{tot} . Let $P_{\mathbf{x}_A^{IS}} = \text{tr}(\mathbf{Q}_{\mathbf{x}_A^{IS}})$ and $P_{\mathbf{x}_A^{AN}} = \text{tr}(\mathbf{Q}_{\mathbf{x}_A^{AN}})$, which satisfies $P_{\mathbf{x}_A} = P_{\mathbf{x}_A^{IS}} + P_{\mathbf{x}_A^{AN}} \leq P_A^{tot}$. Also, the jamming power of $\mathbf{x}_E$ satisfies $P_{\mathbf{x}_E} = \text{tr}(\mathbf{Q}_{\mathbf{x}_E}) \leq P_E^{tot}$, where P_E^{tot} is the total power constraint of Eve. We assume that the transmitted signals on different streams are uncorrelated, so that $\mathbf{Q}_{\mathbf{x}_A^{IS}}$, $\mathbf{Q}_{\mathbf{x}_A^{AN}}$, and $\mathbf{Q}_{\mathbf{x}_E}$ are diagonal positive semidefinite matrices.

B. CSI Assumption

Although most of existing works assume that the global channel information is available at Alice and Eve, e.g., [8], [14], we consider a more practical CSI assumption. In our model, both Alice and Eve only know partial channel state information. Specifically, we consider a system that operates in the time division duplex (TDD) model. Due to the reciprocity of the channel in TDD systems, we assume that Alice knows $\sqrt{g_{AB}}\mathbf{H}_{AB}$ by utilizing pilots from Bob during the training phase, instead of the feedback from Bob, so that Eve can hardly intercept this channel information. Therefore, Eve is only assumed to know $\sqrt{\rho}\mathbf{H}_{EE}$, $\sqrt{g_{AE}}\mathbf{H}_{AE}$ and $\sqrt{g_{EB}}\mathbf{H}_{EB}$, while has no information of $\sqrt{g_{AB}}\mathbf{H}_{AB}$. The self-loop channel $\mathbf{H}_{EE}$ can be estimated beforehand by Eve, and $\sqrt{g_{AE}}\mathbf{H}_{AE}$ and $\sqrt{g_{EB}}\mathbf{H}_{EB}$ could be estimated by overhearing the pilot signals sent by Alice and Bob, respectively. Besides, we suppose that Bob can estimate $\mathbf{K}_B$ in the last TS and feedback this information to Alice. Similarly, Eve is supposed to know the $\mathbf{S}_E$ and $\mathbf{K}_E$ in the last TS.

Since it is assumed that Alice does not know $\mathbf{H}_{AE}$, similar to [50], Alice uniformly allocates the transmit power of AN to each signal in $\mathbf{x}_A^{AN}$, such that $\mathbf{Q}_{\mathbf{x}_A^{AN}} = \frac{P_{\mathbf{x}_A^{AN}}}{N_A - N_B}\mathbf{I}_{N_A - N_B}$. Therefore, the equation (8) can be rewritten as

$$\begin{aligned} \mathbf{K}_E &= \frac{g_{AE}P_{\mathbf{x}_A^{AN}}}{N_A - N_B}\mathbf{H}_{AE}\mathbf{W}_A^{AN}(\mathbf{W}_A^{AN})^H\mathbf{H}_{AE}^H \\ &\quad + \rho\mathbf{H}_{EE}\mathbf{Q}_{\mathbf{x}_E}\mathbf{H}_{EE}^H + \sigma_E^2\mathbf{I}_{N_E}. \end{aligned} \quad (9)$$

III. PROBLEM FORMULATION

In this section, we first give some necessary backgrounds about the stochastic game for self-contained. Then, considering the interactions between Eve and Alice, we formulate the eavesdropping problem in the MIMO system as an imperfect-information stochastic game.

A. Background Concepts

1) Imperfect-Information Stochastic Game [36], [51]:

An N -agent stochastic game with imperfect information can be regarded as the extension of the partial observable Markov decision process (POMDP) in multi-agent scenarios. It can be described by a tuple $\Gamma = \langle \mathcal{S}, \mathcal{A}^1, \dots, \mathcal{A}^N, r^1, \dots, r^N, \mathcal{Z}^1, \dots, \mathcal{Z}^N, O, p, \gamma \rangle$. At each time step $t \in \{0, 1, \dots, \infty\}$, $s_t \in \mathcal{S}$ describes the state of the environment, and each agent $n \in \mathcal{N} = \{1, 2, \dots, N\}$ executes an action $a_t^n \in \mathcal{A}^n$. Actions of all agents form a joint action $\mathbf{a}_t \in \mathcal{A} = \mathcal{A}^1 \times \dots \times \mathcal{A}^N$. This incurs a transition in the environment according to the state transition function $p(s_{t+1}|s_t, \mathbf{a}_t)$ (which typically is unknown): $\mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0, 1]$. The agent n receives the immediate reward $r_t^n(s_t, \mathbf{a}_t)$ (abbreviated to r_t^n below) as a consequence of taking the previous actions. $\gamma \in [0, 1]$ is the reward discount factor across time.

At the time-step t , each agent n chooses actions based on the current state s_t using the private policy $\pi^n \in \Pi^n$, which is the probability density function conditioned on the state, i.e., $a_t^n \sim \pi^n(\cdot|s_t)$, and Π^n is the policy space for agent n . We denote the policy profile as $\pi = \{\pi^1, \dots, \pi^N\}$. The standard joint value function (V-function) V_n^π and the action-value function (Q-function) Q_n^π are respectively defined as

$$\begin{aligned} V_n^\pi(s_t) &= \mathbb{E}_{\tau \sim \pi} \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k}^n | s_t \right\}, \\ Q_n^\pi(s_t, \mathbf{a}_t) &= \mathbb{E}_{\tau \sim \pi} \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k}^n | s_t, \mathbf{a}_t \right\}, \end{aligned} \quad (10)$$

where $\mathbb{E}_{\tau \sim \pi} \{\cdot\}$ means that the expectation is taken over the trajectory $\tau = (s_t, \mathbf{a}_t, s_{t+1}, \mathbf{a}_{t+1}, \dots)$, which is generated by executing the policy profile π . Let $p_0(s)$ denote the distribution of the initiate state s_0 , the expected performance objective for each agent n is given by

$$J_n(\pi) = \mathbb{E}_{s \sim p_0} \{V_n^\pi(s)\}. \quad (11)$$

Assume that the system dynamic is ergodic, and let $\mu_\pi(s)$ be the steady state distribution under the policy profile π , i.e., $\mu_\pi(s) = \lim_{t \rightarrow \infty} \Pr(s_t = s | \pi)$. According to the chapter 10.4 in [24], given that $p_0(s) = \mu_\pi(s)$,² the objective (11) is proportional to the average reward, i.e.,

$$J_n(\pi) = \frac{1}{1-\gamma} \mathbb{E}_{\tau \sim \pi} \left\{ \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=0}^T r_t^n \right\} = \frac{1}{1-\gamma} \mathbb{E}_t \{r_t^n | \pi\}, \quad (12)$$

where $\mathbb{E}_t \{r_t^n | \pi\}$ means the expectation taken over time steps t when using the policy profile π . Therefore, maximizing the discounted objective (11) is essentially equivalent to maximizing the average reward.

In a partial observable scenario, each agent can only obtain the individual observation according to observation function $O(s, \mathbf{a}, n) : \mathcal{S} \times \mathcal{A}^n \times \mathcal{N} \rightarrow \mathcal{Z}^n$. Therefore, each agent n can only choose actions based on their local observation, i.e., $a_t^n \sim \pi^n(\cdot|z_t^n)$.

²This could be easily realized in the practical implementations. For example, let the systems run a long while using the given policy profile π and then randomly select a state, the sampled state obeys the distribution μ_π .

2) *Best Response*: Let π^{-n} refer to the set of all policies in π other than π^n . Given a fixed strategy profile π^{-n} , the best responses of the agent n against π^{-n} is defined as

$$\beta(\pi^{-n}) = \arg \max_{\pi^n} J_n(\pi^n, \pi^{-n}). \quad (13)$$

3) *Nash Equilibrium* [32]: A Nash equilibrium of a stochastic game is a policy profile in which each agent's policy is the best response against the policies of the others. Formally, a policy profile $\pi_* = \{\pi_*^1, \dots, \pi_*^N\}$ is a Nash equilibrium if for all $n \in \mathcal{N}$, it satisfies

$$J_n(\pi_*) \geq J_n(\pi^n, \pi_*^{-n}), \forall \pi^n. \quad (14)$$

4) *Rational* [29]: Given that the policies of other agents are stable, the learning algorithm is said to be rational if the agent's policy using this algorithm converges to the best response to other agents.

5) *Convergent* [29]: In a multi-agent stochastic game, a learning algorithm is convergent if the agent's policy using this algorithm will necessarily converge to a stationary state. This convergence property is usually affected by the learning algorithms used by other agents. In general, existing works focused on the convergence in **self-play** (i.e., all of the agents use the same learning algorithm). Therefore, the algorithm is said to be convergent if all agents' policies converge to a stationary state in the self-play.

6) *Relationship to Equilibrium*: In the multi-agent game, if all agents use a rational algorithm and this algorithm converges finally, the convergence state must be a Nash equilibrium [29]. This conclusion can be proved by using the definitions of rational, convergent, and Nash equilibrium. This provides a criterion for judging whether the empirical results learned by algorithms are Nash equilibrium.

B. Stochastic Game Representation

In this subsection, we formulate the eavesdropping at Eve and anti-eavesdropping at Alice as a stochastic game with imperfect information. In our setting, there are only two agents, i.e., $\mathcal{N} = \{1, 2\}$, so that the game is described as $\Gamma = \langle \mathcal{S}, \mathcal{A}^1, \mathcal{A}^2, r^1, r^2, \mathcal{Z}^1, \mathcal{Z}^2, O, p, \gamma \rangle$. Without loss of generality, we let Alice be the agent 1 and Eve be the agent 2. For clarity, in the rest of this paper, we replace the integer superscripts 1 and 2 with the capital letters 'A' and 'E', respectively. Therefore, the game is re-denoted as $\Gamma = \langle \mathcal{S}, \mathcal{A}^A, \mathcal{A}^E, r^A, r^E, \mathcal{Z}^A, \mathcal{Z}^E, O, p, \gamma \rangle$.

1) *Action*: Let $\mathcal{D}_k^+$ denote the set of all diagonal positive semidefinite matrices with size k . According to the descriptions in Sec. II-A, the action space of Eve is formally denoted by $\mathcal{A}^E = \{\mathbf{Q} | \mathbf{Q} \in \mathcal{D}_{N_E}^+, \text{tr}(\mathbf{Q}) \leq P_E^{\text{tot}}\}$. The action at the time-step t is defined as

$$a_t^E = \mathbf{Q}_{\mathbf{x}_E}(t) \in \mathcal{A}^E. \quad (15)$$

As for Alice, since the power of AN signal is distributed uniformly, the action space is $\mathcal{A}^A = \{(\mathbf{Q}, P) | \mathbf{Q} \in \mathcal{D}_{N_B}^+, \text{tr}(\mathbf{Q}) + P \leq P_A^{\text{tot}}\}$, and the action at the time-step t is

$$a_t^A = (\mathbf{Q}_{\mathbf{x}_A^{\text{IS}}}(t), P_{\mathbf{x}_A^{\text{AN}}}(t)) \in \mathcal{A}^A. \quad (16)$$

The joint action is

$$\mathbf{a}_t = (a_t^E, a_t^A). \quad (17)$$

2) *State and Observation*: At the time-step t , we assume that the system state $s_t \in \mathcal{S}$ is the global channel information of all links. That is,

$$s_t = (\sqrt{g_{AB}}\mathbf{H}_{AB}(t), \sqrt{g_{AE}}\mathbf{H}_{AE}(t), \sqrt{g_{EB}}\mathbf{H}_{EB}(t), \sqrt{\rho}\mathbf{H}_{EE}(t)). \quad (18)$$

However, as discussed in Sec. II-B, both Alice and Eve only have the partial information: Alice only knows $\sqrt{g_{AB}}\mathbf{H}_{AB}(t)$, and Eve knows $\sqrt{g_{AE}}\mathbf{H}_{AE}(t)$, $\sqrt{g_{EB}}\mathbf{H}_{EB}(t)$, $\sqrt{\rho}\mathbf{H}_{EE}(t)$. Besides, Alice knows the previous signal and IN covariances, $(\mathbf{S}_B(t-1), \mathbf{K}_B(t-1))$, and Eve knows $(\mathbf{S}_E(t-1), \mathbf{K}_E(t-1))$. Since the channels vary slowly in adjacent time slots, the previous signal and IN covariances also imply some information of the current channel state. In order to help the learning algorithm to better adjust the policy in the absence of the global channel information, we introduce the previous signal and IN covariances into the observation. Therefore, at the time-step t , the local observation of Alice $z_t^A \in \mathcal{Z}^A$ is given by

$$z_t^A = (\sqrt{g_{AB}}\mathbf{H}_{AB}(t), \mathbf{S}_B(t-1), \mathbf{K}_B(t-1)), \quad (19)$$

and the local observation of Eve $z_t^E \in \mathcal{Z}^E$ is given by

$$z_t^E = (\sqrt{g_{AE}}\mathbf{H}_{AE}(t), \sqrt{g_{EB}}\mathbf{H}_{EB}(t), \sqrt{\rho}\mathbf{H}_{EE}(t), \mathbf{S}_E(t-1), \mathbf{K}_E(t-1)). \quad (20)$$

Note that although the transition of the channel state is irrelevant with the action, due to the incorporating of the previous signal and IN covariances into the local observation, the actions taken by Alice and Eve will influence the following observation, and then influence the future rewards and actions in the dynamic learning process. Besides, we will also study the impact of partial channel state information on the performance of both sides in the simulation part.

3) *Reward*: At time-step t , the channel capacities of Bob and Eve are respectively given by [52]

$$C_B(t) = \log(\det(\mathbf{I}_{N_B} + \mathbf{S}_B(t)\mathbf{K}_B^{-1}(t))), \quad (21)$$

and

$$C_E(t) = \log(\det(\mathbf{I}_{N_E} + \mathbf{S}_E(t)\mathbf{K}_E^{-1}(t))). \quad (22)$$

According to [53], when both Alice and Bob have knowledge of the Eve's channel, Alice can adopt the secrecy-capacity achieving scheme so that Eve is unable to decode any fixed fraction of the message bits. However, in this paper, we consider a more practical scenario, in which neither Alice nor Bob know the Eve's channel. As a result, Alice adopts the coding scheme with a rate of $C_B(t)$ to transmit information bits to Bob. According to [7] and [13], if $C_E(t) \geq C_B(t)$, Eve can reliably decode the information bits, otherwise it can not decode the information at an arbitrarily small probability of error. Since the logarithmic function is a monotonically increasing function of $\det(\cdot)$, to make the reward (to be defined later) more sensitive to the action variation and thus accelerate the learning speed, we replace $C_B(t)$ and $C_E(t)$ with $\tilde{C}_B(t) = \det(\mathbf{I}_{N_B} + \mathbf{S}_B(t)\mathbf{K}_B^{-1}(t))$ and $\tilde{C}_E(t) = \det(\mathbf{I}_{N_E} + \mathbf{S}_E(t)\mathbf{K}_E^{-1}(t))$, respectively, in the following reward function design.

As for Eve, it aims to reliably eavesdrop on as much information as possible. Therefore, the reward function of Eve is designed as

$$r^E(s_t, \mathbf{a}_t) = \tilde{C}_B(t) \cdot \mathbb{1}\{\tilde{C}_E(t) \geq \tilde{C}_B(t)\}, \quad (23)$$

where the indicator function $\mathbb{1}\{x\} = 1$ if the boolean value x is True, and 0 otherwise.

As for Alice, it aims to transmit as much information as possible securely. If $\tilde{C}_B(t) \geq \tilde{C}_E(t)$, information security is guaranteed. Otherwise, the information is leaked, and there will be penalties for each leaked bit. The reward function of the Alice (or the suspicious Alice-Bob pair) is designed as

$$r^A(s_t, \mathbf{a}_t) = \tilde{C}_B(t)(1 - \zeta_A^i \mathbb{1}\{\tilde{C}_E(t) \geq \tilde{C}_B(t)\}), \quad (24)$$

where ζ_A^i is the penalty factor of information leakage, and it can be adjusted according to the required level of confidentiality. For example, if the information is insignificant, ζ_A^i can be set to 0, and for very important information, we can increase ζ_A^i to guide the learning algorithm to focus on the confidentiality.

It should be noticed that, in real-world scenarios, it is difficult for the suspicious pair to know whether the information is leaked, i.e., the leakage indicator $\mathbb{1}\{C_E(t) \geq C_B(t)\}$ is typically unavailable at Alice or Bob. However, similar to the worst-case scenario considered in [21], we set a strong adversary for the legitimate eavesdropper by assuming that Alice knows the leakage indicator. This setup helps Eve achieve better performance in practical oppositional communication environments since it is difficult for Alice to obtain the leakage indicator information for anti-wiretap.

C. Problem Formulation

In the formulated game Γ , the target of Alice is

$$\max_{\pi^A} J_A(\pi^A, \pi^E). \quad (25)$$

According to the discussions in Sec. III-A, the discounted objective is proportional to the average reward, and because the logarithmic function is monotonic, the problem (25) is equivalent to

$$\max_{\pi^A} \mathbb{E}_t\{C_B(t)(1 - \zeta_A^i \mathbb{1}\{C_E(t) \geq C_B(t)\}) | \pi^A, \pi^E\}. \quad (26)$$

The target of Eve is

$$\max_{\pi^E} J_E(\pi^A, \pi^E), \quad (27)$$

which is equivalent to

$$\max_{\pi^E} \mathbb{E}_t\{C_B(t) \mathbb{1}\{C_E(t) \geq C_B(t)\} | \pi^A, \pi^E\}. \quad (28)$$

The objective function in (28) is termed as **eavesdropping rate** [7]. Two other important metrics in eavesdropping tasks are the **relative eavesdropping rate** and **eavesdropping non-outage probability**, which are defined as [7]

$$\frac{\mathbb{E}_t\{C_B(t) \mathbb{1}\{C_E(t) \geq C_B(t)\} | \pi^A, \pi^E\}}{\mathbb{E}_t\{C_B(t) | \pi^A, \pi^E\}}, \quad (29)$$

and

$$\mathbb{E}_t\{\mathbb{1}\{C_E(t) \geq C_B(t)\} | \pi^A, \pi^E\}, \quad (30)$$

respectively. We stress that solving (27) does not necessarily maximize the relative eavesdropping rate and eavesdropping non-outage probability, while we use them as demonstrated metrics in the simulation results.

According to the definition of Nash equilibrium, if both of Alice and Eve achieve their goals, the policy profile $\pi = \{\pi^A, \pi^E\}$ will be a Nash equilibrium. According to the theorem 3.1 in [54], we can obtain the following proposition about the existence of Nash equilibrium.

Proposition 1: For each agent $n \in \{A, E\}$, suppose that the policy space Π^n is equicontinuous³ and bounded.⁴ Then there exists at least one Nash equilibrium in the considered problem.

Proof: We give the proof in Appendix A.

IV. LEARNING ALGORITHM FOR STOCHASTIC GAME

The objectives of Alice and Eve are to solve the problems (25) and (27), respectively. For the practical purpose, we assume that Alice and Eve neither know each other's behavior nor exchange information. Therefore, Alice or Eve learns independently using the local trajectories $\{z_t^n, a_t^n, r_t^n, z_{t+1}^n\}_{t \geq 0}$, $n \in \{A, E\}$ collected by interacting with others. In the independent learning framework, the opponent can be treated as part of the environment. Therefore the objective of the agent in the framework could be regarded as a single-agent task

$$\max_{\pi} J(\pi), \quad (31)$$

where the subscript about agent index is omitted since we consider an arbitrary agent. The most straightforward way for solving (31) is to apply a common RL algorithm on each agent to learn the best response. However, the problem is not a real single-agent task, but rather a task in which the opponent policy would continuously change during the learning process. Thereby, the policy learned by common RL methods would be unstable, which may lead to non-convergence and performance degradation.

A. FSP Algorithm

In order to tackle the instability and find the Nash equilibrium in multi-agent tasks, [34] introduced a general framework named Fictitious Self-Play (FSP). The brief pseudo-code is given by Algorithm 1. Since the learning process does not require extra information from other agents, we omit the agent index in the description of the algorithm. In this algorithm, each round j of learning includes three main phases. In the first data generation phase (i.e., GenerateData function), the agent interacts with opponents using the mixture policy σ of the best response β_j and the average policy π_{j-1} . Specifically, at the beginning of the data generation phase, we set the policy as β_j with probability η_j and as π_{j-1} with probability $1 - \eta_j$, where η_j is called 'anticipatory parameter'. Then the agent chooses actions using the mixture policy σ . The collected

data is stored in two separated memory buffers: the memory buffer $\mathcal{M}_{RL}$ for RL stores all the transitions (z_t, a_t, r_t, z_{t+1}) since off-policy RL algorithms can learn the best response from historical experiences, and the memory buffer $\mathcal{M}_{SL}$ for supervised learning stores the agent's behavior from the best response β_j since the average policy π aims to fit the average policy of all the past best responses of its own. In the second phase, the agent updates its best response by RL (i.e., RLProcess function) using the data from $\mathcal{M}_{RL}$. Finally, the agent updates the average strategy by supervised learning (i.e., SLProcess function) using data from $\mathcal{M}_{SL}$.

Algorithm 1 [34] General FSP for an Arbitrary Agent

function FSP(Γ)

Initialize $j = 1$, policy π_{j-1} , and $\beta_j \leftarrow \pi_{j-1}$

Initialize memory buffer $\mathcal{M}_{SL}, \mathcal{M}_{RL}$

for $j = 1, 2, \dots$, until convergence **do**

Choose anticipatory parameter η_j

$\mathcal{M}_{SL}, \mathcal{M}_{RL} \leftarrow \text{GenerateData}(\pi_{j-1}, \beta_j, \eta_j)$

$\beta_{j+1} \leftarrow \text{RLProcess}(\mathcal{M}_{RL})$

$\pi_j \leftarrow \text{SLProcess}(\mathcal{M}_{SL})$

end for

return π_j

end function

function GenerateData(π, β, η)

Set policy $\sigma = \begin{cases} \pi, & \text{with probability } 1 - \eta \\ \beta, & \text{with probability } \eta, \end{cases}$

Initialize $t = 1$ and observation z_1

for $t = 1, 2, \dots, T$ **do**

Execute action a_t using policy σ

Observe reward r_t and next observation z_{t+1}

Store all transition (z_t, a_t, r_t, z_{t+1}) into $\mathcal{M}_{RL}$

if agent choose action from policy β **then**

Store (z_t, a_t) into $\mathcal{M}_{SL}$

end if

return $\mathcal{M}_{SL}, \mathcal{M}_{RL}$

end function

The original FSP adopted a table-based RL and supervised learning algorithm, while in our problem, the state and action spaces are continuous, so that the table-based FSP can not be utilized straightforwardly. Reference [33] proposed the NFSP-DQN by combining FSP with DQN, and NFSP-DQN approached the Nash equilibriums in imperfect information games with high-dimensional observation space. However, DQN can only solve problems with low-dimensional and discrete action space. Although we can discretize the action, DQN performs badly in discrete action space with a large number of feasible actions, and the action discretization would bring extra quantization errors. Besides, due to DQN can only choose action deterministically, NFSP-DQN can not solve games with Mixed strategy Nash equilibrium (MSNE), even very simple games such as Rock-Paper-Scissor.

In order to deal with the continuous observation and action spaces in our problem, we develop the NFSP-SAC by combining the FSP with SAC, which is a deep RL algorithm based on the actor-critic framework. In NFSP-SAC, both the best response policy β and average policy π are parameterized by

³For every $\epsilon > 0$, there exists $\delta > 0$, such that if $|a_i^n - a_j^n| < \delta$, $a_i^n, a_j^n \in \mathcal{A}^n$, then $|\pi^n(a_i|s) - \pi^n(a_j|s)| < \epsilon, \forall \pi^n \in \Pi^n, s \in \mathcal{S}$.

⁴There exists a finite number $M > 0$ such that $\max_{s \in \mathcal{S}, a^n \in \mathcal{A}^n} |\pi^n(a^n|s)| \leq M$.

artificial neural networks (NN). We will give the RL process and supervised learning process in detail in the following two subsections.

B. Reinforcement Learning Stage

In this subsection, we use SAC to learn the best response β using the historical data stored in $\mathcal{M}_{RL}$. Different from traditional RL methods, SAC augments the objective function by adding the entropy of the policy to the reward. By doing so, the exploration of the policy is enhanced, and the agent is encouraged to search several near-optimal policies, avoiding falling into the local optimum in the game. With a current policy π , the entropy augmented reward is given by [35], [55],

$$r_t^\pi = r_t + \alpha \mathcal{H}(\pi(\cdot|s_t)), \quad (32)$$

where $\mathcal{H}(p) = -\int_{\mathcal{X}} p(x) \log p(x) dx$ is the differential entropy function, and α is the temperature parameter which indicates the relative importance of the entropy in the reward. Therefore, the standard value function is transferred to the following entropy augmented one,

$$V^\pi(s_t) = \mathbb{E}_{\tau \sim \pi} \left\{ \sum_{k=0}^{\infty} \gamma^k [r_{t+k} + \alpha \mathcal{H}(\pi(\cdot|s_{t+k}))] | s_t \right\}, \quad (33)$$

which is called soft V-function. Note that superscript is π rather than the joint policy π since we learn from the view of single agent. Similarly, the soft Q-function is given by

$$\begin{aligned} Q^\pi(s_t, a_t) &= \mathbb{E}_{\tau \sim \pi} \left\{ \sum_{k=0}^{\infty} \gamma^k r_{t+k} + \alpha \sum_{k=1}^{\infty} \gamma^k \mathcal{H}(\pi(\cdot|s_{t+k})) \right\} | s_t, a_t \\ &= r_t + \gamma \mathbb{E}_{s_{t+1} \sim p} \{ V^\pi(s_{t+1}) \}. \end{aligned} \quad (34)$$

The relationship between Q^π and V^π is given by the following derivation

$$\begin{aligned} V^\pi(s_t) &= \mathbb{E}_{a_t \sim \pi} \{ Q^\pi(s_t, a_t) \} + \alpha \mathcal{H}(\pi(\cdot|s_t)) \\ &= \mathbb{E}_{a_t \sim \pi} \{ Q^\pi(s_t, a_t) \} - \alpha \int_{\mathcal{A}} \pi(a|s_t) \log \pi(a|s_t) da \\ &= \mathbb{E}_{a_t \sim \pi} \{ Q^\pi(s_t, a_t) - \alpha \log \pi(a_t|s_t) \}. \end{aligned} \quad (35)$$

Therefore, the objective is to find the optimal policy which maximizes the following expected soft V-function,

$$\begin{aligned} J(\pi) &= \mathbb{E}_{s \sim p_0} \{ V^\pi(s) \} \\ &= \mathbb{E}_{s \sim p_0} \{ \mathbb{E}_{a \sim \pi} \{ Q^\pi(s, a) - \alpha \log \pi(a|s) \} \}. \end{aligned} \quad (36)$$

When α approaches to 0, the objective in (36) degenerates to the standard form in (25) or (27).

Because action spaces of both Alice and Eve are continuous in our setting, the value-based RL methods, such as DQN, are not suitable for solving this problem. Therefore, we will use the policy-based RL method to maximize the objective defined in (36). In the policy-based RL methods, the policy is updated by gradient ascent algorithms with the policy gradient [24], i.e., the gradient of the objective with respect to the policy. In practical implementations of DRL, the policy is typically parameterized to a tractable function using NN. In this work, we parameterize π to π_θ with parameters θ . With the parameterization of the policy π , the policy gradient of $J(\pi_\theta)$ is given by the following theorem.

Theorem 1: Let $p(s \rightarrow s', t, \pi_\theta)$ denote the state transition probability from s to s' after t steps with policy π_θ . The discounted state distribution (as in [56]) is defined as $\rho_{\pi_\theta}(s') = (1-\gamma) \int_{s \in \mathcal{S}} \sum_{t=0}^{\infty} \gamma^t p_0(s) p(s \rightarrow s', t, \pi_\theta) ds$. Then,

$$\begin{aligned} \nabla_\theta J(\pi_\theta) &= \frac{1}{1-\gamma} \mathbb{E}_{s \sim \rho_{\pi_\theta}, a \sim \pi_\theta} \{ \nabla_\theta \log \pi_\theta(a|s) (Q^{\pi_\theta}(s, a) \\ &\quad - V^{\pi_\theta}(s) - \alpha \log \pi_\theta(a|s)) \}, \end{aligned} \quad (37)$$

where $\mathbb{E}_{s \sim \rho_{\pi_\theta}, a \sim \pi_\theta} \{ \cdots \} = \int_{\mathcal{S}} \rho_{\pi_\theta}(s) \int_{\mathcal{A}} \pi_\theta(a|s) \{ \cdots \} da ds$.

Proof: We provide the proof in Appendix B.

In practical implementations, the agents will calculate the gradient using the off-line data stored in $\mathcal{M}_{RL}$, while the distribution of these states does not obey $\rho_{\pi_\theta}(s)$ in general, which may introduce biases. However, according to the Off-policy Policy-gradient theorem in [46], replacing $\rho_{\pi_\theta}(s)$ with an arbitrary distribution $\rho_b(s)$ in (37) is also a good gradient to improve the policy. Therefore, by ignoring the constant coefficient and replacing ρ_π with ρ_b , the policy gradient (37) is modified to

$$\begin{aligned} \Delta \theta &= \mathbb{E}_{s \sim \rho_b, a \sim \pi_\theta} \{ \nabla_\theta \log \pi_\theta(a|s) \\ &\quad \times (Q^{\pi_\theta}(s, a) - V^{\pi_\theta}(s) - \alpha \log \pi_\theta(a|s)) \}. \end{aligned} \quad (38)$$

Updating the policy by gradient ascent with (38), the policy provably converges to at least a local optimum under some strict conditions [46]. However, these conditions for guaranteeing the convergence are so strict that the learning speed is slow and thus impractical. Based on the policy gradient in (38), we will give the practical algorithm. In DRL, the soft Q-function Q^π is represented by Q_ω with parameters ω , and soft V-function V^π is represented by V_ϕ with parameters ϕ . In principle, we do not need a separate approximator V_ϕ for V^π , due to the relationship between V^π and Q^π in Eq. (35), but in practice, this extra function approximator can stabilize the learning process [35]. Replacing Q^π with Q_ω , the estimation of $\nabla_\theta J(\pi_\theta)$ with one sample z_t drawn from $\mathcal{M}_{RL}$ is given by

$$\begin{aligned} \Delta \theta_t &= \nabla_\theta \log \pi_\theta(a_t|z_t) (Q_\omega(z_t, a_t) \\ &\quad - V_\phi(z_t) - \alpha \log \pi_\theta(a_t|z_t)) \end{aligned} \quad (39)$$

where we substitute state s_t in policy and value functions with z_t , since the agent only has the partial observation in our partial observable setting. It should be noticed that, to ensure that $\Delta \theta_t$ in (39) is an unbiased estimation of $\Delta \theta$ in (38), the action a_t in (39) is generated according to $\pi_\theta(\cdot|z_t)$, rather than sampled from $\mathcal{M}_{RL}$.

Compared with updating the gradient using one sample, it is usually more stable and computationally advantageous to update the parameters by using mini-batch data, i.e., a part of the sampled trajectory. Consequently, the mini-batch update for θ using the gradient ascent is given by

$$\theta \leftarrow \theta + \eta_\theta \sum_{i=1}^B \Delta \theta_i, \quad (40)$$

where η_θ is the step size of θ , i is the renumbered sample tag, and B is the size of mini-batch data.

The initial value functions are parameterized arbitrarily, which should be updated to the accurate values. The parameters ϕ for V_ϕ are optimized by minimizing the following the mean-squared error (MSE) loss function,

$$L(\phi) = \frac{1}{2} \mathbb{E}_{z_t \sim \mathcal{M}_{RL}} \left\{ (V_\phi(z_t) - \mathbb{E}_{a_t \sim \pi_\theta} [Q_\omega(z_t, a_t) - \alpha \log \pi_\theta(a_t|z_t)])^2 \right\}, \quad (41)$$

where $\mathbb{E}_{z_t \sim \mathcal{M}_{RL}} \{\dots\}$ means the sample average over z_t sampled from $\mathcal{M}_{RL}$ uniformly. The unbiased gradient estimation of the loss in (41) is given by

$$\Delta\phi_t = \nabla_\phi V_\phi(z_t) (V_\phi(z_t) - Q_\omega(z_t, a_t) + \alpha \log \pi_\theta(a_t|z_t)). \quad (42)$$

In (42), a_t is generated based on $\pi_\theta(\cdot|z_t)$ to maintain the unbiasedness. The parameters ω for Q_ω are optimized by minimizing the squared Temporal-Difference (TD) error,

$$L(\omega) = \mathbb{E}_{(z_t, a_t, r_t, z_{t+1}) \sim \mathcal{M}_{RL}} \left\{ \frac{1}{2} (Q_\omega(z_t, a_t) - \hat{Q}(z_t, a_t))^2 \right\}, \quad (43)$$

where

$$\hat{Q}(z_t, a_t) = r_t + \gamma \mathbb{E}_{z_{t+1} \sim p(\cdot|z_t, a_t)} \{V_{\bar{\phi}}(z_{t+1})\}. \quad (44)$$

Similar to DQN, we adopt a target V-function $V_{\bar{\phi}}$ to stabilize the learning process, whose parameters $\bar{\phi}$ are updated by moving-average. That is, in each training step, $\bar{\phi}$ is updated according to,

$$\bar{\phi} \leftarrow \nu \phi + (1 - \nu) \bar{\phi}, \quad (45)$$

where $\nu \in (0, 1]$ is the moving average parameter. In the sampled fashion, the gradient approximator of (43) is given by

$$\Delta\omega_t = \nabla_\omega Q_\omega(z_t, a_t) (Q_\omega(z_t, a_t) - r_t - \gamma V_{\bar{\phi}}(z_{t+1})), \quad (46)$$

where (z_t, a_t, r_t, z_{t+1}) is drawn from $\mathcal{M}_{RL}$. Then, the parameters of V_ϕ and Q_ω are updated by using mini-batch gradient descent,

$$\phi \leftarrow \phi - \eta_\phi \sum_{i=1}^B \Delta\phi_i, \quad (47)$$

$$\omega \leftarrow \omega - \eta_\omega \sum_{i=1}^B \Delta\omega_i, \quad (48)$$

where η_ϕ and η_ω are the respective step sizes of ϕ and ω .

C. Supervised Learning of NFSP-SAC

Let π_{avg} be the real unknown average policy of the previous best responses, and the approximated average policy is parameterized by π_ψ . In the supervised learning stage, our objective is to fit the approximated average policy to the real one. The similarity between two distributions can be measured by the Kullback-Leibler (KL) divergence. For example, letting p, q be two probability density functions defined over x , the KL divergence between p and q is

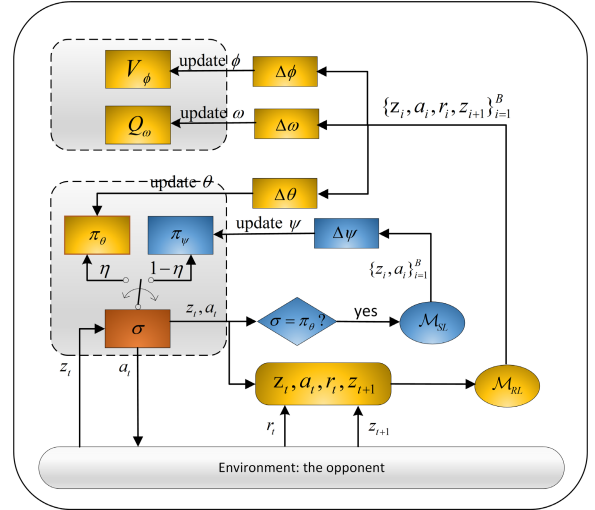


Fig. 2. Overview of the structure of NFSP-SAC.

$D_{KL}(p||q) = \int_x p(x) \log \frac{p(x)}{q(x)} dx$. Therefore, we fit π_ψ to π_{avg} by minimizing the following expected KL divergence,

$$\begin{aligned} L(\psi) &= \mathbb{E}_{z_t} \{ D_{KL}(\pi_{avg}(\cdot|z_t) || \pi_\psi(\cdot|z_t)) \} \\ &= \mathbb{E}_{z_t} \left\{ \int_{a_t} \pi_{avg}(a_t|z_t) \log \frac{\pi_{avg}(a_t|z_t)}{\pi_\psi(a_t|z_t)} da_t \right\} \\ &= \mathbb{E}_{z_t; a_t \sim \pi_{avg}} \{ \log \pi_{avg}(a_t|z_t) - \log \pi_\psi(a_t|z_t) \} \end{aligned} \quad (49)$$

With a given data (z_t, a_t) drawn from $\mathcal{M}_{SL}$, the unbiased gradient estimation of Eq. (45) is given by

$$\Delta\psi_t = -\nabla_\psi \log \pi_\psi(a_t|z_t), \quad (50)$$

which is independent of π_{avg} . The update rule for θ_ψ within a given mini-batch is

$$\psi \leftarrow \psi - \eta_\psi \sum_{i=1}^B \Delta\psi_i. \quad (51)$$

D. NFSP-SAC Algorithm

The structure of NFSP-SAC for an arbitrary agent is depicted in Fig. 2, where the parts of RL stage are painted with yellow, and the parts of SL stage are blue. Besides, the detailed pseudo-code is given in Algorithm 2. At the beginning of each episode, the agent selects the policy π_θ or π_ψ according to the anticipatory parameter η , and the selected policy is used in this episode. Buffer $\mathcal{M}_{SL}$ stores $(z_t, a_t)_{t>0}$ only when $\sigma = \pi_\theta$. $\mathcal{M}_{RL}$ stores all the transitions $(z_t, a_t, r_t, z_{t+1})_{t>0}$. In each training step, the agent uniformly draws data from $\mathcal{M}_{RL}$, and then updates parameters θ , ϕ and ω according to (40), (47) and (48), respectively. Parameters of the approximated average policy ψ are updated according to (51) using data from $\mathcal{M}_{SL}$.

In the practical implementations, we set $\mathcal{M}_{RL}$ as a buffer with a fixed size, where the data storage is subject to the first-in-first-out principle. This is because in the multi-agent game, the opponent's policy is continuously changing and thus the agent can not learn the best response to the current policy of the opponent using the data sampled from too earlier. The SL buffer $\mathcal{M}_{SL}$ is a reservoir buffer that stores all the input data since the averaged policy aims to intimate all the historical

behaviors from the best response policy. Besides, different from the original FSP in Algorithm 1, we carry out the data collection and gradient update alternately, since running both of the RL and SL processes until completely convergence would be computationally expensive.

Algorithm 2 NFSP-SAC for an Arbitrary Agent

Initiate value functions Q_ω, V_ϕ with arbitrary parameters ω, ϕ
 Initiate target function $V_{\bar{\phi}}$ by assigning ϕ to $\bar{\phi}$
 Initiate policies π_θ, π_ψ with parameters θ, ψ
 Initialize memory buffers $\mathcal{M}_{RL}$ and $\mathcal{M}_{SL}$
 Set global step counter $C_{step} = 0$
 Set anticipatory parameter η
 Set update interval N_u be a integer
for episode = 1, 2, $\dots$ **until** terminal **do**
 Set policy $\sigma = \begin{cases} \pi_\psi, & \text{with probability } 1 - \eta \\ \pi_\theta, & \text{with probability } \eta, \end{cases}$
 Observe the initial observation z_1
 for each step $t = 1, 2, \dots, T$ **do**
 Agent executes action using policy σ
 Get the reward r_t , and the next observation z_{t+1}
 Update the global step counter $c_{step} \leftarrow C_{step} + 1$
 Store the transition (z_t, a_t, r_t, z_{t+1}) into $\mathcal{M}_{RL}$
 if the agent choose actions according to π_θ **then**
 Store the state-action pair (z_t, a_t) into $\mathcal{M}_{SL}$
 if C_{step} is divided by N_u evenly **do**
 Sample data from $\mathcal{M}_{RL}$ and reindex $\{z_i, a_i, r_i, z_{i+1}\}_{i=1}^B$
 Update $\Delta\theta$ according to (40)
 Update ϕ and ω according to (47) and (48) respectively
 Update target function: $\bar{\phi} \leftarrow \tau\phi + (1 - \tau)\bar{\phi}$
 Sample data from $\mathcal{M}_{SL}$ and reindex $\{z_i, a_i\}_{i=1}^B$
 Update ψ according to (51)
 end if
 end for
end for

V. NUMERICAL SIMULATIONS

In this section, we first give the simulation settings. Then we provide the simulation results with discussion and analysis.

A. Simulation Setups

1) *System Setups*: We will use the following parameters without other specified. We set $N_A = 4$ and $N_B = N_E^r = N_E^t = 2$. The noise power is assumed to be the same at Eve and Bob: $\sigma_B^2 = \sigma_E^2 = 10^{-8} \text{ mW}$. The power constraints of Alice and Eve are $P_A^{tot} = 3 \text{ mW}$ and $P_E^{tot} = 10 \text{ mW}$, respectively. The self-interference coefficient $\rho = 5 \times 10^{-10}$. We fix the distance between Alice and Bob to $d_{AB} = 200 \text{ m}$, and the path-loss exponent $\lambda = 3.48$. Eve moves along the perpendicular bisector of the line between Alice and Bob, so that $d_{AE} = d_{EB}$. We default $d_{AE} = d_{EB} = 200 \text{ m}$, and the impact brought by the distance variation would be investigated. The channel transition is modeled as a Markov process. That is, $\mathbf{H}(t+1) = \sqrt{1-\kappa}\mathbf{H}(t) + \sqrt{\kappa}\Phi(t)$, where $\mathbf{H} \in \{\mathbf{H}_{AB}, \mathbf{H}_{AE}, \mathbf{H}_{EB}\}$, the random variable $\Phi(t)$ obeys the

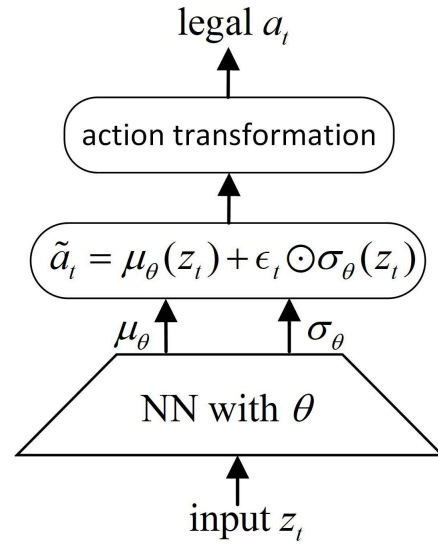
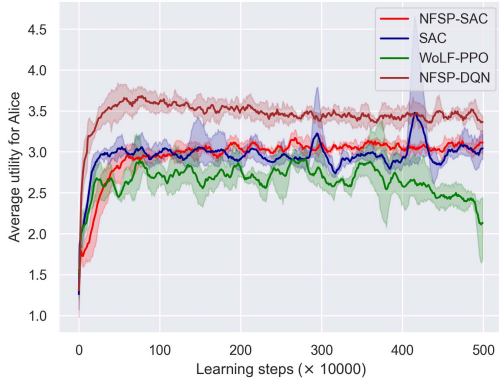


Fig. 3. The structure of policy π_θ .

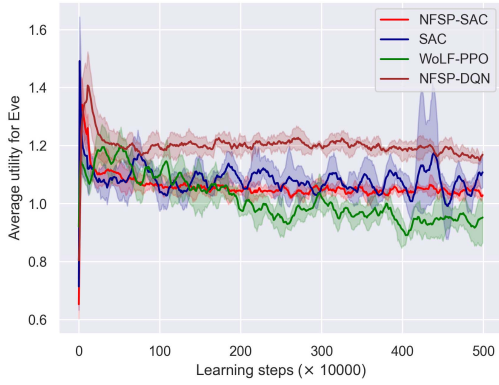
complex Gaussian distribution $\mathcal{CN}(0, 1)$, and $\kappa = 0.05$. The imaginary channel $\mathbf{H}_{EE}$ varies slowly, so that we simplify it to be a constant matrix. The penalty factor of information leakage $\zeta_A^i = 2$.

2) *Structure of Policy*: The detailed structure of policy is given in Fig. 3 (taking π_θ for example). Given the input z_t , NN outputs the means $\mu_\theta(z_t)$ and standard deviations $\sigma_\theta(z_t)$. The raw action $\tilde{a}_t = \mu_\theta(z_t) + \epsilon_t \odot \sigma_\theta(z_t)$, where ϵ_t obeys the standard Gaussian distribution, and $\odot$ denotes the element-wise product. Therefore, the raw action $\tilde{a}_t$ is a non-bounded Gaussian variable. To satisfy the power constraints, we transform the raw action according to the following method, i.e., action transformation in Fig. 3. We define a function $c(x) = \frac{\tanh(x)+1}{2}$, where $\tanh(x)$ is the hyperbolic tangent function. This function maps x into the interval $[0, 1]$. Let $\tilde{a}_t^E = (\tilde{a}_t^E(1), \dots, \tilde{a}_t^E(N_E^t + 1))$ represent the raw action of Eve and $\text{diag}[\dots]$ denote the diagonal matrix. Then transmit power of Eve is $P_{\mathbf{x}_E}(t) = c(\tilde{a}_t^E(1)) \cdot P_E^{tot}$, and the covariance matrix of the signal is $\mathbf{Q}_{\mathbf{x}_E}(t) = P_{\mathbf{x}_E}(t) \cdot \frac{\text{diag}[c(\tilde{a}_t^E(2)), \dots, c(\tilde{a}_t^E(N_E^t + 1))]}{\sum_{i=2}^{N_E^t+1} c(\tilde{a}_t^E(i))}$. For Alice, the raw action is a vector $\tilde{a}_t^A = (\tilde{a}_t^A(1), \dots, \tilde{a}_t^A(N_B + 2))$. We let $P_{\mathbf{x}_A}(t) = c(\tilde{a}_t^A(1))P_A^{tot}$. Then $P_{\mathbf{x}_A^{IS}}(t) = c(\tilde{a}_t^A(2)) \cdot P_{\mathbf{x}_A}(t)$, $\mathbf{Q}_{\mathbf{x}_A^{IS}}(t) = P_{\mathbf{x}_A^{IS}}(t) \cdot \frac{\text{diag}[c(\tilde{a}_t^A(3)), \dots, c(\tilde{a}_t^A(N_B + 2))]}{\sum_{i=3}^{N_B+2} c(\tilde{a}_t^A(i))}$, and $P_{\mathbf{x}_A^{AN}}(t) = P_{\mathbf{x}_A}(t) - P_{\mathbf{x}_A^{IS}}(t)$. By the above action transformation, the peak power constraints can be satisfied. Moreover, raw actions are stored in $\mathcal{M}_{RL}$ and $\mathcal{M}_{SL}$ during the learning process.

3) *Hyperparameters*: We set the learning rates $\eta_\theta = \eta_\psi = \eta_\omega = \eta_\phi = 3 \times 10^{-4}$, discount factor of rewards $\gamma = 0.99$, the temperature parameter $\alpha = 0.05$, and the anticipatory parameter $\eta = 0.1$ [33]. All the policies and value functions are parameterized by adopting the two-layer Rectified Linear Unit (ReLU) Multi-Layer Perceptron (MLP) with 128 units per layer. The length of each episode $T = 1000$. For each interval $N_u = 100$, parameters are updated for 10 rounds. The mini-batch size $B = 128$, and the moving average factor is set to



(a) The learning curve of Alice



(b) The learning curve of Eve

Fig. 4. Comparison of learning processes among different algorithms.

be $\nu = 0.005$. The sizes of memories are 200 thousand and 2 million for $\mathcal{M}_{RL}$ and $\mathcal{M}_{SL}$ respectively.

We will demonstrate the simulation results by comparing our work with following benchmark algorithms:

- 1) SAC: We apply this common RL algorithm to both Alice and Eve. Each agent makes decisions only according to the policy learned by SAC. The parameter update rules have been given in Sec IV-B.
- 2) WoLF-PPO [36]: This method combined the WoLF principle [29] with proximal policy optimization (PPO) algorithm [57]. When the policy is better than the averaged one, the agent learns conservatively by adopting a low learning rate, otherwise, it would update the policy parameters aggressively using a high learning rate. The WoLF principle maintains an average policy π_{avg} . If

$$\int_A \pi_\theta(a|s_t) Q^\pi(s_t, a) ds \geq \int_A \pi_{avg}(a|s_t) Q^\pi(s_t, a) ds, \quad (52)$$

the learning rate $\eta_\theta = \eta_l$, otherwise $\eta_\theta = \eta_h$, where $\frac{\eta_h}{\eta_l} > 1$. Since the integration (52) is hard to calculate, we discrete the action for simplicity. Specifically, each action dim is discrete to 10 intervals. For more details please refer to [29], [36] and [57].

- 3) NFSP-DQN [33]: NFSP-DQN is the combination of FSP and DQN. The main difference from our NFSP-SAC is that DQN is adopted in the RL phase. NFSP-DQN can approach the Nash equilibriums in two poker games.

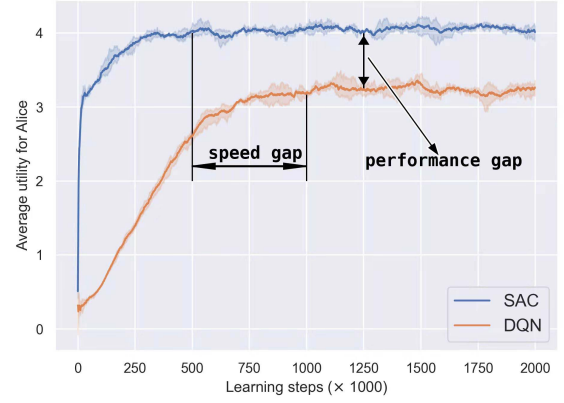
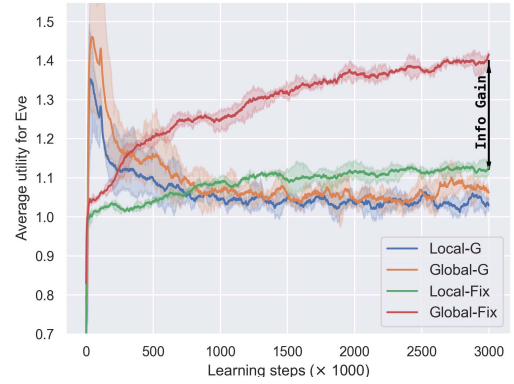
Fig. 5. The utility performance comparisons between DQN and SAC at Alice's side with fixed P_{x_E} .

Fig. 6. The impact of available channel information on eavesdropping performance when confronting the fixed or smart Alice.

This method is only suitable for problems with low dimensional and discrete action space. Therefore, similar to Wolf-PPO, we discrete the actions in NFSP-DQN for comparisons.

B. Simulation Results

Fig. 4 compares the average utility of different approaches in the same eavesdropping and anti-eavesdropping system. For each algorithm, we run four independent experiments. The shadows are the fluctuation ranges during the learning processes. It is shown that our NFSP-SAC converges to a steady-state asymptotically with a small fluctuation range. According to the relationship of convergence and rationality to equilibrium stated at the end of Sec. III-A, since NSFP-SAC is rational which is inherited from SAC, the convergence state is a Nash equilibrium. However, SAC fluctuates severely in the whole learning process and has several sharp peaks. This is because the continuous change of the opponent's policy will lead to the change of the system dynamics, such that the fundamental requirements of convergence of common RL algorithms are violated. Although WoLF-PPO has converged to the Nash equilibrium on several simple matrix games, e.g., the Prisoner's Dilemma and Matching Pennies games, it performs badly in our task, which is more complicated. The non-convergence of WoLF-based methods on complicated tasks also was observed in [58]. Note that the learning process

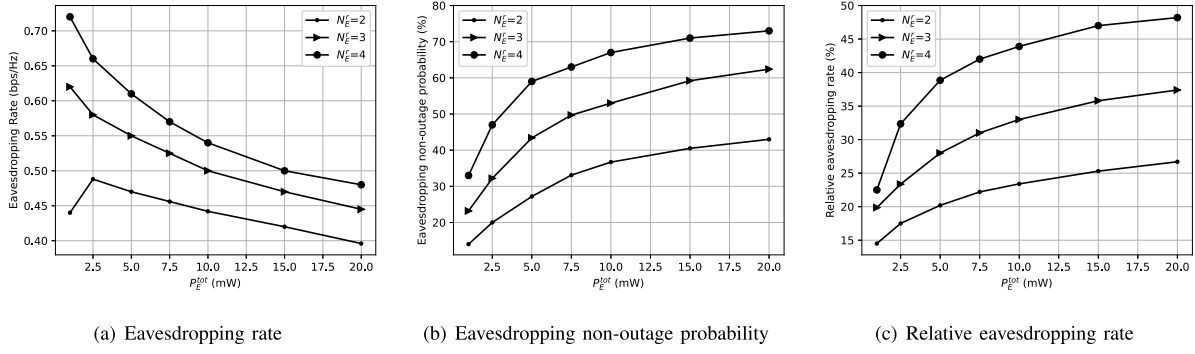


Fig. 7. The performances versus P_E^{tot} with different N_E^r , where (a) is the eavesdropping rate, (b) is the eavesdropping non-outage probability, and (c) is the relative eavesdropping rate.

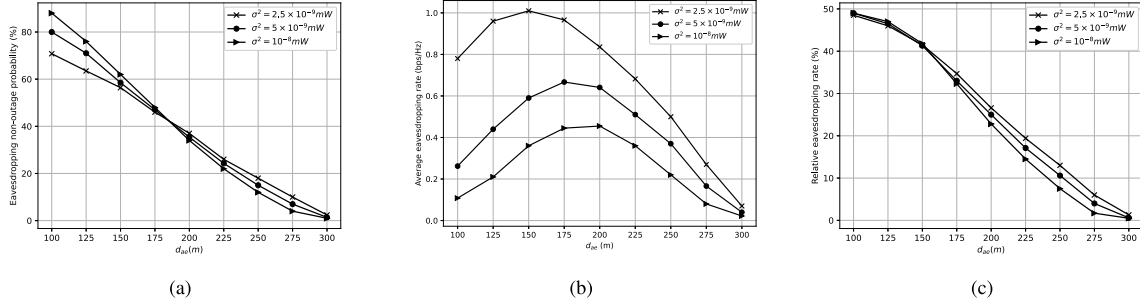


Fig. 8. (a) Eavesdropping non-outage probability, (b) Eavesdropping rate and (c) Relative eavesdropping rate versus different d_{AE} with different noise power.

of NFSP-DQN also asymptotically converges to a stable state, and the average utilities of both Alice and Eve are even higher than that of NFSP-SAC. However, although DQN is rational in theory, it lacks rationality in problems with large discrete actions space. Therefore, this result is far away from the Nash equilibrium. An intuitive result is that Alice should transmit all the feasible power, i.e., $P_A(t) = 3mW, \forall t$. However, we find that the average transmit power of Alice in NFSP-DQN is only $2.36mW$, while in NFSP-SAC it is about $2.98mW$. Actually, in games, the utilities at Nash equilibrium are not necessarily the highest. A well-known example is the Prisoner's Dilemma, in which the term of penalty for both two prisoners at the Nash equilibrium is longer than that at one of the non-equilibrium points. To further verify that the DQN is lacking in rationality, and thus the policy profile learned by NFSP-DQN is not a Nash equilibrium, we make the following comparison.

In Fig. 5, the power covariance matrix of Eve $\mathbf{Q}_{x_E}$ is fixed to $\frac{1}{4}P_E^{tot}\mathbf{I}_{N_E^t}$ for jamming, and both SAC and DQN are applied to Alice. It is shown that DQN is inferior to SAC in both learning speed and performance. In particular, DQN takes one million steps to achieve the steady-state, while for SAC, it only takes about a second as many steps as DQN. Besides, there is a gap of about 0.8 between the steady-state utilities learned by SAC and DQN, which verifies that DQN is lacking in rationality in the considered problem. The degradation in performance of DQN may result from the quantization of action space. As described in the structure of policy in Sec. V-A, the dimension of $\tilde{a}_t^E$ is $N_b + 2$. In our settings with $N_b = 2$, we have discretized each element in a_t^E with 10 intervals, then there are $11^{N_b+2} = 14641$ available actions, which significantly increases the difficulty of learning. The comparison results in Fig. 5 imply that in the RL stage of

FSP, the policy learned by DQN is not the best response to the opponent. Therefore, the policy profile learned by NFSP-DQN is further away from the Nash equilibrium than that learned by NFSP-SAC.

In Fig. 6, we investigate the impact of partial channel information on eavesdropping performance. The observation at Alice remains the same as that defined in Eq. (19). The term 'Local' denotes that only the local observation (defined in Eq. (20)) is available at Eve, and 'Global' means that Eve knows all the channel information, including $\mathbf{H}_{AB}$. The suffix '-Fix' denotes that Alice applies a fixed power for sending information signal and AN. In particular, we fix that $\mathbf{Q}_{x_A^{IS}} = 0.55P_A^{tot}\frac{\mathbf{I}_{N_B}}{2}$, and $P_{x_A^{AN}} = 0.45P_A^{tot}$. The suffix '-G' indicates the game formulated in Sec. III. Each curve is plotted by running four independent experiments. When Alice adopts the fixed action, the problem is degenerated to a regular single-agent RL problem, so that we can train the policy of Eve using SAC. In the game scenario, NFSP-SAC is adopted. It is shown that the average utility of Global-Fix significantly outperforms that of Local-Fix. This indicates that when Alice's policy is fixed, knowing $\mathbf{H}_{AB}$ would bring great benefits to Eve. However, the gap of steady-states between local-G and global-G is marginal, which means that knowing more channel information may not bring more utility to Eve in the game. A possible reason is that although Eve has more information to adjust its policy, Alice also can adapt its policy to reduce transmission power against Eve's eavesdropping.

In Fig. 7, we have illustrated the relationships of eavesdropping rate, non-outage probability, and the relative eavesdropping rate and different constraint of power at Eve P_E^{tot} . Moreover, the effect of the receiving antennas of Eve N_E^r on the eavesdropping performance is studied. The definitions

of these metrics can be found in Sec. III-C. In Fig. 7a, except for the case with $P_E^{tot} = 1$ and $N_E^r = 2$, the eavesdropping rate decreases with the increase of P_E^{tot} . This is because when the jamming power increases, the intelligent Alice also could reduce information signal power $P_{x_{aIS}}$ and enhances the AN power $P_{x_{aAN}}$ to combat the eavesdropping. The decreasing eavesdropping rate in the game does not mean that we should use a small P_E^{tot} in game scenarios, since both the eavesdropping non-outage probability and the relative eavesdropping rate significantly increase with the increasing of P_E^{tot} . Obviously, the average reward learned by Alice should be greater than or equal to 0, since the worst case for Alice is not to send any information signals. Therefore, for the well-trained policy profile $\{\pi^A, \pi^E\}$, it is conceivable that $\mathbb{E}_t\{C_B(t)(1 - \zeta_A^i \mathbb{1}\{C_E(t) \geq C_B(t)\})|\pi^A, \pi^E\} \geq 0 \Leftrightarrow \frac{C_B(t)\mathbb{1}\{C_E(t) \geq C_B(t)|\pi^A, \pi^E\}}{\mathbb{E}_t\{C_B(t)|\pi^A, \pi^E\}} \leq \frac{1}{\zeta_A^i}$. The right hand at the two-way arrow is the relative eavesdropping rate. Therefore, in the current setting with $\zeta_A^i = 2$, the upper bound of the relative eavesdropping rate is 50%. Note that When $N_E^r = 4$ and $P_E^{tot} = 20$, the relative eavesdropping rate is very close to this bound.

In Fig. 8, we present the eavesdropping non-outage probability, eavesdropping rate and relative eavesdropping rate versus the distance between Eve and Alice with different noise power ($\sigma_B^2 = \sigma_E^2 = \sigma^2$). As shown in Fig. 8a and 8c, with the increasing of d_{AE} , both the eavesdropping non-outage probability and relative eavesdropping rate decrease due to the increase of path-loss. Besides, it is shown that the decline rate of both the eavesdropping non-outage probability and relative eavesdropping rate with distance d_{AE} slightly increase with the increasing of noise power. When d_{AE} increases to 300m, the eavesdropping non-outage probability approaches to zero, which implies that in the high path loss scenario, e.g., urban area, it may be difficult to eavesdrop on information even if proactive eavesdropping is taken. For the eavesdropping rate, as shown in Fig 8b, it does not increase monotonically with the decreasing d_{AE} . Because as d_{AE} decreases, the quality of the eavesdropping channel becomes better than that of the channel between Alice and Bob. When Alice finds it difficult to avoid being monitored, it would reduce $P_{x_{aIS}}$, so that the eavesdropping rate decreases. Also, we find that when $\sigma^2 = 10^{-8}mW$, the peak eavesdropping rate is achieved at $d_{AE} = 200m$, and when the noise power decreases, d_{AE} corresponding to the peak eavesdropping rate also reduces.

VI. CONCLUSION

In this paper, we have investigated an adversarial scenario in a multiple-input multiple-output (MIMO) system between a legitimate eavesdropper and a suspicious communication pair comprised of a transmitter and a receiver. Due to the lack of channel information and dynamic variation, this adversarial problem is formulated as a stochastic game with partial observation. To obtain the Nash equilibrium, a deep reinforcement learning based algorithm, NFSP-SAC, is developed by combining FSP with the state-of-art RL method SAC. The simulation results demonstrate that NFSP-SAC can converge to a steady state, overcoming the fluctuation of common RL methods in the game. Since NFSP-SAC is rational, according

to the relationship of convergence and rationality, the steady state is a Nash equilibrium. Besides, by changing the power penalty factor, the available CSI, and the location of Eve, it is observed that transmitting more jamming power, having more CSI, and better quality of the eavesdropping channel does not necessarily lead to higher eavesdropping rate as Alice can intelligently adjust its policy to combat the fiercely attacking.

APPENDIX A PROOF OF PROPOSITION 1

We give the proof in N -agent scenarios, which is more general.

According to the theorem 3.1 in [54], if Π^n (policy space of agent n) is compact and convex, $J_n(\cdot, \pi^{-n})$ is quasi-concave on Π^n , and J_n is continuous on $\Pi = \times_{n \in \mathcal{N}} \Pi^n$, then there exists at least one Nash equilibrium point $\pi_* = \{\pi_*^1, \dots, \pi_*^N\}$. Utilizing the content of mathematical analysis in [59], it is easy to prove that if Π^n is equicontinuous and bounded, then it is compact and convex. In the following Lemma 1 and Lemma 2, we will prove that $J_n(\cdot, \pi^{-n})$ is linear (also quasi-concave) on Π^n and J_n is continuous on Π , respectively.

Define the distance between any two functions f and g as

$$d(f, g) = \max_{s \in \mathcal{S}, a \in \mathcal{A}} |f(a|s) - g(a|s)|. \quad (53)$$

The distance defined in Eq. (53) is obviously a metric.

Lemma 1: In the considered system, for any agent n , a policy $\pi^n = \alpha\pi_1^n + (1 - \alpha)\pi_2^n, \forall \alpha \in [0, 1], \forall \pi_1^n, \pi_2^n \in \Pi^n$, and every $\pi^{-n} \in \Pi^{-n}$, it holds that

$$J_n(\pi^n, \pi^{-n}) = \alpha J_n(\pi_1^n, \pi^{-n}) + (1 - \alpha) J_n(\pi_2^n, \pi^{-n}). \quad (54)$$

Proof: As in [56], let $p(s \rightarrow s', t, \pi)$ denote the state transition probability from s to s' after t steps with policy π . The discounted state distribution is defined as $\rho_\pi(s) = (1 - \gamma) \int_{\mathcal{S}} \sum_{t=0}^{\infty} \gamma^t p_0(s_0) p(s_0 \rightarrow s, t, \pi) ds_0$. According to [56],

$$\begin{aligned} J_n(\pi^n, \pi^{-n}) &= \frac{1}{1 - \gamma} \int_{\mathcal{S}} \rho_\pi(s) \int_{\mathcal{A}} \pi(a|s) r^n(s, a) da ds \\ &= \frac{1}{1 - \gamma} \int_{\mathcal{S}} \rho(s) \int_{\mathcal{A}^n} \pi^n(a|s) \bar{r}^n(s, a^n) da^n ds \end{aligned} \quad (55)$$

where $\bar{r}^n(s, a^n) = \int_{\mathcal{A}^{-n}} \pi^{-n}(a^{-n}|s) r^n(s, a) da^{-n}$, and $\rho_\pi(s)$ is changed to $\rho(s)$ since the channel state transition in our system has no relevant with the policies. Then plugging $\pi^n = \alpha\pi_1^n + (1 - \alpha)\pi_2^n$ into Eq. (55), it is readily obtained that

$$J_n(\pi) = \alpha J_n(\pi_1^n, \pi^{-n}) + (1 - \alpha) J_n(\pi_2^n, \pi^{-n}). \quad (56)$$

This completes the proof. $\square$

Lemma 2: For each $n \in \mathcal{N}$, J_n is continuous on Π .

Proof: For $\pi, \pi_0 \in \Pi$, let $A_n^{\pi_0}(s, a) = Q_n^{\pi_0}(s, a) - V_n^{\pi_0}(s)$. According to [60], we have

$$\begin{aligned} |J_n(\pi) - J_n(\pi_0)| &= \left| \frac{1}{1 - \gamma} \int_{\mathcal{S}} \rho_\pi(s) \int_{\mathcal{A}} \pi(a|s) A_n^{\pi_0}(s, a) da ds \right| \\ &\leq \frac{1}{1 - \gamma} \int_{\mathcal{S}} \rho_\pi(s) \left| \int_{\mathcal{A}} \pi(a|s) A_n^{\pi_0}(s, a) da \right| ds \end{aligned} \quad (57)$$

It was proved in [60] that $|\int_{\mathcal{A}} \pi(\mathbf{a}|s) A_n^{\pi_0}(s, \mathbf{a}) d\mathbf{a}| \leq 2 d_{TV}(\pi, \pi_0) \max_{s, \mathbf{a}} |A_n^{\pi_0}(s, \mathbf{a})|$, where

$$\begin{aligned} d_{TV}(\pi, \pi_0) &= \frac{1}{2} \max_{s \in \mathcal{S}} \int_{\mathcal{A}} |\pi(\mathbf{a}|s) - \pi_0(\mathbf{a}|s)| d\mathbf{a} \\ &\leq \frac{\max_{s \in \mathcal{S}, \mathbf{a} \in \mathcal{A}} |\pi(\mathbf{a}|s) - \pi_0(\mathbf{a}|s)|}{2} \int_{\mathcal{A}} d\mathbf{a} \\ &= \frac{d(\pi, \pi_0)}{2} \int_{\mathcal{A}} d\mathbf{a}, \end{aligned} \quad (58)$$

Since the action space $\mathcal{A}$ is bounded, $M_{\mathcal{A}} = \int_{\mathcal{A}} d\mathbf{a}$ is a finite number. Therefore, we have

$$|J_n(\pi) - J_n(\pi_0)| < \frac{d(\pi, \pi_0) M_{\mathcal{A}} \max_{s, \mathbf{a}} |A_n^{\pi_0}(s, \mathbf{a})|}{1 - \gamma}. \quad (59)$$

For any $\epsilon > 0$, let $\delta = \frac{\epsilon(1-\gamma)}{M_{\mathcal{A}} \max_{s, \mathbf{a}} |A_n^{\pi_0}(s, \mathbf{a})|}$. Then whenever $d(\pi, \pi_0) < \delta$, $|J_n(\pi) - J_n(\pi_0)| < \epsilon$ holds. This completes the proof. $\square$

APPENDIX B PROOF OF THEOREM 1

This proof follows the same lines of the standard stochastic policy gradient theorem in [61]. We start from the derivative of $V^{\pi_\theta}(s)$,

$$\begin{aligned} \nabla_\theta V^{\pi_\theta}(s) &= \nabla_\theta \int_{\mathcal{A}} \left(\pi_\theta(a|s) (Q^{\pi_\theta}(s, a) - \alpha \log \pi_\theta(a|s)) \right) d\mathbf{a} \\ &\stackrel{(a)}{=} \int_{\mathcal{A}} \left\{ \nabla_\theta \pi_\theta(a|s) Q^{\pi_\theta}(s, a) \right. \\ &\quad \left. + \pi_\theta(a|s) \nabla_\theta (r(s, a) + \gamma \int_{\mathcal{S}} p(s' \rightarrow s', 1, \pi_\theta) V^{\pi_\theta}(s') ds') \right. \\ &\quad \left. - \alpha \nabla_\theta \pi_\theta(a|s) \log \pi_\theta(a|s) - \alpha \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s) \right\} d\mathbf{a} \\ &\stackrel{(b)}{=} \int_{\mathcal{A}} \nabla_\theta \pi_\theta(a|s) (Q^{\pi_\theta}(s, a) - \alpha \log \pi_\theta(a|s)) d\mathbf{a} \\ &\quad + \gamma \int_{\mathcal{A}} \pi_\theta(a|s) \int_{\mathcal{S}} p(s' \rightarrow s', 1, \pi_\theta) \nabla_\theta V^{\pi_\theta}(s') ds' d\mathbf{a} \\ &= \int_{\mathcal{A}} \nabla_\theta \pi_\theta(a|s) (Q^{\pi_\theta}(s, a) - \alpha \log \pi_\theta(a|s)) d\mathbf{a} \\ &\quad + \gamma \int_{\mathcal{S}} p(s \rightarrow s', 1, \pi_\theta) \nabla_\theta V^{\pi_\theta}(s') ds', \end{aligned} \quad (60)$$

where in (a), the last term is omitted since $\int_{\mathcal{A}} \pi_\theta(a|s) \nabla_\theta \log \pi_\theta(a|s) d\mathbf{a} = \int_{\mathcal{A}} \nabla_\theta \pi_\theta(a|s) d\mathbf{a} = 0$, and in the second term of (b), we exchange the order of integration according to Fubini's theorem, and then substituting $\int_{\mathcal{A}} \pi_\theta(a|s) p(s' \rightarrow s', 1, \pi_\theta) d\mathbf{a}$ with $p(s \rightarrow s', 1, \pi_\theta)$. Then, unrolling the above formula iteratively, we have

$$\begin{aligned} \nabla_\theta V^{\pi_\theta}(s) &= \int_{\mathcal{A}} \nabla_\theta \pi_\theta(a|s) (Q^{\pi_\theta}(s, a) - \alpha \log \pi_\theta(a|s)) d\mathbf{a} \\ &\quad + \gamma \int_{\mathcal{S}} p(s \rightarrow s', 1, \pi_\theta) \\ &\quad \int_{\mathcal{A}} \nabla_\theta \pi_\theta(a|s') (Q^{\pi_\theta}(s', a) - \alpha \log \pi_\theta(a|s')) d\mathbf{a} ds' \end{aligned}$$

$$\begin{aligned} &+ \gamma^2 \int_{\mathcal{S}} p(s \rightarrow s', 1, \pi_\theta) \\ &\quad \times \int_{\mathcal{S}} p(s' \rightarrow s'', 1, \pi_\theta) \nabla_\theta V^{\pi_\theta}(s'') ds'' ds' \\ &= \int_{\mathcal{A}} \nabla_\theta \pi_\theta(a|s) (Q^{\pi_\theta}(s, a) - \alpha \log \pi_\theta(a|s)) d\mathbf{a} \\ &\quad + \gamma \int_{\mathcal{S}} p(s \rightarrow s', 1, \pi_\theta) \\ &\quad \times \int_{\mathcal{A}} \nabla_\theta \pi_\theta(a|s') (Q^{\pi_\theta}(s', a) - \alpha \log \pi_\theta(a|s')) d\mathbf{a} ds' \\ &\quad + \gamma^2 \int_{\mathcal{S}} p(s \rightarrow s', 2, \pi_\theta) \nabla_\theta V^{\pi_\theta}(s') ds' \\ &\stackrel{(c)}{=} \int_{\mathcal{S}} \sum_{t=0}^{\infty} \gamma^t p(s \rightarrow s', t, \pi_\theta) \\ &\quad \times \int_{\mathcal{A}} \nabla_\theta \pi_\theta(a|s') (Q^{\pi_\theta}(s', a) - \alpha \log \pi_\theta(a|s')) d\mathbf{a} ds', \end{aligned} \quad (61)$$

where (c) holds because $\lim_{t \rightarrow \infty} \gamma^t \int_{\mathcal{S}} p(s \rightarrow s', t, \pi_\theta) \nabla_\theta V^{\pi_\theta}(s') ds' = 0$. Then, taking the expectation over the initial state, we have

$$\begin{aligned} \nabla_\theta J(\pi_\theta) &= \int_{\mathcal{S}} p_0(s) \nabla_\theta V^{\pi_\theta}(s) ds \\ &= \int_{\mathcal{S}} p_0(s) \int_{\mathcal{S}} \sum_{t=0}^{\infty} \gamma^t p(s \rightarrow s', t, \pi_\theta) \\ &\quad \times \int_{\mathcal{A}} \nabla_\theta \pi_\theta(a|s') (Q^{\pi_\theta}(s', a) - \alpha \log \pi_\theta(a|s')) d\mathbf{a} ds' ds \\ &= \frac{1}{1-\gamma} \int_{\mathcal{S}} \rho_\pi(s') \int_{\mathcal{A}} \pi_\theta(a|s') \\ &\quad \times \frac{\nabla_\theta \pi_\theta(a|s')}{\pi_\theta(a|s')} (Q^{\pi_\theta}(s', a) - \alpha \log \pi_\theta(a|s')) d\mathbf{a} ds' \\ &\stackrel{(d)}{=} \frac{1}{1-\gamma} \mathbb{E}_{s \sim \rho_\pi, a \sim \pi_\theta} \{ \nabla_\theta \log \pi_\theta(a|s) \\ &\quad \times (Q^{\pi_\theta}(s, a) - V^{\pi_\theta}(s) - \alpha \log \pi_\theta(a|s)) \}, \end{aligned} \quad (62)$$

where the term $V^{\pi_\theta}(s)$ in (d) is the baseline to reduce the variance in practical implementations, and it does not introduce any bias due to the following derivation,

$$\begin{aligned} \mathbb{E}_{s \sim \rho_\pi, a \sim \pi_\theta} \{ \nabla_\theta \log \pi_\theta(a|s) V^{\pi_\theta}(s) \} &= \int_{\mathcal{S}} \rho_\pi(s) \int_{\mathcal{A}} \nabla_\theta \pi_\theta(a|s) V^{\pi_\theta}(s) d\mathbf{a} ds \\ &= \int_{\mathcal{S}} \rho_\pi(s) \nabla_\theta \left(\int_{\mathcal{A}} \pi_\theta(a|s) d\mathbf{a} \right) V^{\pi_\theta}(s) ds = 0. \end{aligned} \quad (63)$$

$\square$

REFERENCES

- [1] Y.-S. Shiu, S. Y. Chang, H.-C. Wu, S. C.-H. Huang, and H.-H. Chen, "Physical layer security in wireless networks: A tutorial," *IEEE Wireless Commun.*, vol. 18, no. 2, pp. 66–74, Apr. 2011.
- [2] S. Goel and R. Negi, "Guaranteeing secrecy using artificial noise," *IEEE Trans. Wireless Commun.*, vol. 7, no. 6, pp. 2180–2189, Jun. 2008.
- [3] G. Zheng, I. Krikidis, J. Li, A. P. Petropulu, and B. Ottersten, "Improving physical layer secrecy using full-duplex jamming receivers," *IEEE Trans. Signal Process.*, vol. 61, no. 20, pp. 4962–4974, Oct. 2013.

- [4] S.-H. Tsai and H. V. Poor, "Power allocation for artificial-noise secure MIMO precoding systems," *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 5983–5993, Jun. 2011.
- [5] D. Xu, J. Cheng, and C. Zhang, "Performance analysis on anti-jamming technique of smart antennas," *Command Inf. Syst. Technol.*, vol. 2, no. 2, pp. 1–5, 2011.
- [6] D. Wang, B. Bai, W. Zhao, and Z. Han, "A survey of optimization approaches for wireless physical layer security," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 2, pp. 1878–1911, 2nd Quart., 2019.
- [7] J. Xu, L. Duan, and R. Zhang, "Proactive eavesdropping via cognitive jamming in fading channels," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2790–2806, May 2017.
- [8] 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.
- [9] D. Hu, Q. Zhang, P. Yang, and J. Qin, "Proactive monitoring via jamming in amplify-and-forward relay networks," *IEEE Signal Process. Lett.*, vol. 24, no. 11, pp. 1714–1718, Nov. 2017.
- [10] 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.
- [11] 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.
- [12] X. Jiang, H. Lin, C. Zhong, X. Chen, and Z. Zhang, "Proactive eavesdropping in relaying systems," *IEEE Signal Process. Lett.*, vol. 24, no. 6, pp. 917–921, Jun. 2017.
- [13] C. Zhong, X. Jiang, F. Qu, and Z. Zhang, "Multi-antenna wireless legitimate surveillance systems: Design and performance analysis," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4585–4599, Jul. 2017.
- [14] 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. 2017.
- [15] J. Moon, H. Lee, C. Song, S. Kang, and I. Lee, "Relay-assisted proactive eavesdropping with cooperative jamming and spoofing," *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6958–6971, Oct. 2018.
- [16] X. Zhou, B. Maham, and A. Hjørungnes, "Pilot contamination for active eavesdropping," *IEEE Trans. Wireless Commun.*, vol. 11, no. 3, pp. 903–907, Mar. 2012.
- [17] J. Jose, A. Ashikhmin, T. L. Marzetta, and S. Vishwanath, "Pilot contamination and precoding in multi-cell TDD systems," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 2640–2651, Aug. 2011.
- [18] A. Mukherjee and A. L. Swindlehurst, "Jamming games in the MIMO wiretap channel with an active eavesdropper," *IEEE Trans. Signal Process.*, vol. 61, no. 1, pp. 82–91, Jan. 2013.
- [19] A. Garnaev, M. Baykal-Gursoy, and H. V. Poor, "A game theoretic analysis of secret and reliable communication with active and passive adversarial modes," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 2155–2163, Mar. 2016.
- [20] A. Garnaev, Y. Liu, and W. Trappe, "Anti-jamming strategy versus a low-power jamming attack when intelligence of adversary's attack type is unknown," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 2, no. 1, pp. 49–56, Mar. 2016.
- [21] 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.
- [22] X. Tang, P. Ren, Y. Wang, and Z. Han, "Combating full-duplex active eavesdropper: A hierarchical game perspective," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1379–1395, Mar. 2017.
- [23] L. Busoni, R. Babuska, and S. B. De, "A comprehensive survey of multiagent reinforcement learning," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 38, no. 2, pp. 156–172, Feb. 2008.
- [24] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.
- [25] L. Xiao, D. Jiang, D. Xu, H. Zhu, Y. Zhang, and H. V. Poor, "Two-dimensional antijamming mobile communication based on reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9499–9512, Oct. 2018.
- [26] A. S. Lakshminarayanan, S. Sharma, and B. Ravindran, "Dynamic action repetition for deep reinforcement learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, San Francisco, CA, USA, Feb. 2017, pp. 2133–2139.
- [27] L. Xiao, C. Xie, M. Min, and W. Zhuang, "User-centric view of unmanned aerial vehicle transmission against smart attacks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3420–3430, Apr. 2018.
- [28] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [29] M. Bowling and M. Veloso, "Multiagent learning using a variable learning rate," *Artif. Intell.*, vol. 136, no. 2, pp. 215–250, Apr. 2002.
- [30] X. He, H. Dai, and P. Ning, "Faster learning and adaptation in security games by exploiting information asymmetry," *IEEE Trans. Signal Process.*, vol. 64, no. 13, pp. 3429–3443, Jul. 2016.
- [31] M. L. Littman, "Markov games as a framework for multi-agent reinforcement learning," in *Proc. 11th Int. Conf. Mach. Learn.*, New Brunswick, NJ, USA, Jul. 1994, pp. 157–163.
- [32] J. Hu and M. P. Wellman, "Nash Q-learning for general-sum stochastic games," *J. Mach. Learn. Res.*, vol. 4, pp. 1039–1069, Nov. 2003.
- [33] J. Heinrich and D. Silver, "Deep reinforcement learning from self-play in imperfect-information games," 2016, *arXiv:1603.01121*.
- [34] J. Heinrich, M. Lanctot, and D. Silver, "Fictitious self-play in extensive-form games," in *Proc. 32nd Int. Conf. Mach. Learn.*, Lille, France, vol. 37, Jul. 2015, pp. 805–813.
- [35] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. 35th Int. Conf. Mach. Learn.*, Stockholm, Sweden, vol. 80, Jul. 2018, pp. 1856–1865.
- [36] D. S. Ratcliffe, K. Hofmann, and S. Devlin, "Win or learn fast proximal policy optimisation," in *Proc. IEEE Conf. Games (CoG)*, London, U.K., Aug. 2019, pp. 1–4.
- [37] N. Yang, S. Yan, J. Yuan, R. Malaney, R. Subramanian, and I. Land, "Artificial noise: Transmission optimization in multi-input single-output wiretap channels," *IEEE Trans. Commun.*, vol. 63, no. 5, pp. 1771–1783, May 2015.
- [38] S. Yan, X. Zhou, N. Yang, B. He, and T. D. Abhayapala, "Artificial-noise-aided secure transmission in wiretap channels with transmitter-side correlation," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8286–8297, Dec. 2016.
- [39] W. Huang, W. Chen, B. Bai, and Z. Han, "Wiretap channel with full-duplex proactive eavesdropper: A game theoretic approach," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 7658–7663, Aug. 2018.
- [40] G. W. Brown, "Iterative solution of games by fictitious play," *Activity Anal. Prod. Allocation*, vol. 13, no. 1, pp. 374–376, 1951.
- [41] L. Xiao, G. Sheng, S. Liu, H. Dai, M. Peng, and J. Song, "Deep reinforcement learning-enabled secure visible light communication against eavesdropping," *IEEE Trans. Commun.*, vol. 67, no. 10, pp. 6994–7005, Oct. 2019.
- [42] R. Hu, "Deep fictitious play for stochastic differential games," *Commun. Math. Sci.*, vol. 19, no. 2, pp. 325–353, 2021.
- [43] D. S. Leslie and E. J. Collins, "Generalised weakened fictitious play," *Games Econ. Behav.*, vol. 56, no. 2, pp. 285–298, Aug. 2006.
- [44] X. Ma, K. Driggs-Campbell, and M. J. Kochenderfer, "Improved robustness and safety for autonomous vehicle control with adversarial reinforcement learning," in *Proc. IEEE Intell. Vehicles Symp.*, Changshu, China, Jun. 2018, pp. 1–7.
- [45] K. Kawamura and Y. Tsuruoka, "Neural fictitious self-play on ELF mini-rts," in *Proc. 33rd AAAI Conf. Artif. Intell.*, Honolulu, HI, USA, Jan. 2019, pp. 1–8.
- [46] T. Degris, M. White, and R. S. Sutton, "Off-policy actor-critic," in *Proc. 29th Int. Conf. Mach. Learn.*, Edinburgh, U.K., Jun. 2012, pp. 1–18.
- [47] V. R. Konda and J. N. Tsitsiklis, "Actor-critic algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, pp. 1008–1014.
- [48] T. Riihonen, S. Werner, and R. Wichman, "Mitigation of loopback self-interference in full-duplex MIMO relays," *IEEE Trans. Signal Process.*, vol. 62, no. 13, pp. 3479–3493, Aug. 2014.
- [49] P. Lioliou, M. Viberg, M. Coldrey, and F. Athley, "Self-interference suppression in full-duplex MIMO relays," in *Proc. Conf. Rec. Forty 4th Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, Nov. 2010, pp. 658–662.
- [50] S. Yan, N. Yang, I. Land, R. Malaney, and J. Yuan, "Three artificial-noise-aided secure transmission schemes in wiretap channels," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3669–3673, Apr. 2017.
- [51] J. N. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proc. AAAI Conf. Artif. Intell.*, New Orleans, LA, USA, Feb. 2018, pp. 2974–2982.
- [52] Z. Wang and G. B. Giannakis, "Outage mutual information of space-time MIMO channels," *IEEE Trans. Inf. Theory*, vol. 50, no. 4, pp. 657–662, Apr. 2004.
- [53] A. Khisti and G. W. Wornell, "Secure transmission with multiple antennas I: The MISOME wiretap channel," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3088–3104, Jul. 2010.
- [54] P. J. Reny, "On the existence of pure and mixed strategy Nash equilibria in discontinuous games," *Econometrica*, vol. 67, no. 5, pp. 1029–1056, 1999.

- [55] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, "Reinforcement learning with deep energy-based policies," in *Proc. 34th Int. Conf. Mach. Learn.*, Sydney, NSW, Australia, vol. 70, Aug. 2017, pp. 1352–1361.
- [56] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. 31st Int. Conf. Mach. Learn.*, Beijing, China, vol. 32, Jun. 2014, pp. 387–395.
- [57] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017, *arXiv:1707.06347*.
- [58] D. Simões, N. Lau, and L. P. Reis, "Mixed-policy asynchronous deep Q-learning," in *Proc. Iberian Robot. Conf.* Cham, Switzerland: Springer, 2017, pp. 129–140.
- [59] (2014). *The Space of Continuous Functions*. [Online]. Available: https://www.math.edu.hk/course_builder/1415/math3060/Chapter%203.%20Continuous%20Functions.pdf
- [60] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1889–1897.
- [61] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, pp. 1057–1063.



Delin Guo received the B.S. degree in electronic information science and technology from Jilin University, Jilin, China, in 2017. He is currently pursuing the Ph.D. degree with the Department of Electronic Science and Engineering, Nanjing University, China. His current research interests include energy harvesting wireless communications, physical-layer security, reinforcement learning, and multiagent reinforcement learning.



Hui Ding received the B.S. and M.S. degrees in signal and information processing from Nanjing University, Nanjing, China, in 2011. He joined the State Key Laboratory of Air Traffic Management System and Technology in 2011. His current research interests include information processing and intelligent decision support system in air traffic management and reinforcement learning.



Lan Tang (Member, IEEE) received the B.S. and M.S. degrees in communication engineering from Jilin University, Jilin, China, in 2002 and 2005, respectively, and the Ph.D. degree in communication and information science from Southeast University, Nanjing, China, in 2009. Since 2009, she has been with the School of Electronic Science and Engineering, Nanjing University, China, where she is currently an Associate Professor. Her current research interests include key technologies in 5G/6G, machine learning and optimization theory in wireless networks, and integrated communication-radar systems.



Xinggan Zhang received the B.S. degree in electrical engineering and the M.S. and Ph.D. degrees from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 1982, 1988, and 2001, respectively. In 1992, he joined the Department of Electronic Engineering at the NUAU, where he was an Associate Professor. In 1999, he joined the Department of Electronic Science and Engineering, Nanjing University, where he is currently a Professor. His research interests include target recognition and image processing.



Lvxi Yang (Senior Member, IEEE) received the M.S. and Ph.D. degrees in electrical engineering from Southeast University, Nanjing, China, in 1990 and 1993, respectively. Since 1993, he has been with the Department of Radio Engineering, Southeast University, where he is currently a Full Professor of information systems and communications, and the Director of the Digital Signal Processing Division. He has authored or coauthored two published books and more than 300 journal articles and holds 60 patents. His current research interests include signal processing for wireless communications, MIMO communications, intelligent wireless communications, and statistical signal processing. He received the First and Second Class Prizes of Science and Technology Progress Awards of the State Education Ministry of China, in 1998, 2002, 2014, and 2018. He is currently a member of Signal Processing Committee of the Chinese Institute of Electronics.



Ying-Chang Liang (Fellow, IEEE) is currently a Professor with the University of Electronic Science and Technology of China, China, where he leads the Center for Intelligent Networking and Communications (CINC). He was a Professor at The University of Sydney, Australia, the Principal Scientist and a Technical Advisor at the Institute for Infocomm Research, Singapore, and a Visiting Scholar at Stanford University, USA. His research interests include wireless networking and communications, cognitive radio, symbiotic communications, dynamic spectrum

access, the internet-of-things, artificial intelligence, and machine learning techniques.

He has been recognized by the Thomson Reuters (now Clarivate Analytics) as a Highly Cited Researcher since 2014. He received the Prestigious Engineering Achievement Award from The Institution of Engineers, Singapore, in 2007, the Outstanding Contribution Appreciation Award from the IEEE Standards Association in 2011, and the Recognition Award from the IEEE Communications Society Technical Committee on Cognitive Networks in 2018. He is the recipient of numerous paper awards, including the IEEE Communications Society Stephen O. Rice Prize in 2021, the IEEE Vehicular Technology Society Jack Neubauer Memorial Award in 2014, and the IEEE Communications Society APB Outstanding Paper Award in 2012. He is a Foreign Member of Academia Europaea. He is the Founding Editor-in-Chief of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS: COGNITIVE RADIO SERIES, and the Key Founder and currently the Editor-in-Chief of the IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING. He is also an Associate Editor-in-Chief of *China Communications*. He was a Guest/Associate Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE JOURNAL OF SELECTED AREAS IN COMMUNICATIONS, the *IEEE Signal Processing Magazine*, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and the IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORK. He was also an Associate Editor-in-Chief of the *World Scientific Journal on Random Matrices: Theory and Applications*. He was a Distinguished Lecturer of the IEEE Communications Society and the IEEE Vehicular Technology Society. He was the Chair of the IEEE Communications Society Technical Committee on Cognitive Networks, and the TPC Chair and Executive Co-Chair of the IEEE Globecom'17.