Reinforcement Learning-based NOMA Power Allocation in the Presence of Smart Jamming

Liang Xiao*, Yanda Li*, Canhuang Dai*, Huaiyu Dai†, H. Vincent Poor‡
* Dept. of Communication Engineering, Xiamen University, China. Email: lxiao@xmu.edu.cn
† Dept. of Electrical and Computer Engineering, NC State University, Raleigh, NC. Email: huaiyu_dai@ncsu.edu
‡ Dept. of Electrical Engineering, Princeton University, Princeton, NJ. Email: poor@princeton.edu

Abstract—Non-orthogonal multiple access (NOMA) systems are vulnerable to jamming attacks, especially smart jammers who apply programmable and smart radio devices such as software defined radios to flexibly control their jamming strategy according to the ongoing NOMA transmission and radio environment. In this paper, the power allocation of a base station in a NOMA system equipped with multiple antennas contending with a smart jammer is formulated as a zero-sum game, in which the base station as the leader first chooses the transmit power on multiple antennas while a jammer as the follower selects the jamming power to interrupt the transmission of the users. A Stackelberg equilibrium of the anti-jamming NOMA transmission game is derived and the conditions assuring its existence are provided to disclose the impact of multiple antennas and radio channel states. A reinforcement learning based power control scheme is proposed for the downlink NOMA transmission without being aware of the jamming and radio channel parameters. The Dyna architecture that formulates a learned world model from the real anti-jamming transmission experience and the hotbooting technique that exploits experiences in similar scenarios to initialize the quality values are used to accelerate the learning speed of the Q-learning based power allocation and thus improve the communication efficiency of NOMA transmission in the presence of smart jammers. Simulation results show that the proposed scheme can significantly increase the sum data rates of users and thus the utilities compared with the standard Q-learning based strategy.

Index Terms—NOMA, smart jamming, power allocation, game theory, reinforcement learning

I. INTRODUCTION

By using power domain multiplexing with successive interference cancellation, non-orthogonal multiple access (NOMA) as an important candidate for 5G cellular communications can significantly improve both the outage performance and user fairness compared with orthogonal multiple access systems [1]. The Multiple-input and multiple-output (MIMO) NOMA transmission can achieve even higher spectral efficiency. For instance, the MIMO NOMA system as proposed in [2] applies zero-forcing based beamforming and pairing to reduce the cluster size of the cellular system and interference and thus significantly increases the capacity. The MIMO NOMA scheme as developed in [3] provides fast transmission of small-size packets following the required outage performance. The MIMO-NOMA system implements successive interference cancellation to detect and remove the signals of the weaker users according to downlink cooperative transmission protocol such as those in [4] and [5].

NOMA transmission is vulnerable to jamming attacks, as an attacker can use programmable and smart radio devices such as software defined radios to flexibly control the jamming power according to the ongoing communication [6]–[9]. As an extreme case, all the downlink users in a NOMA system can be simultaneously blocked, if a jammer sends strong jamming power on the frequency at which the users send their signals. By reducing the signal-to-interference-plus-noise ratio (SINR) of the user signals, jammers can significantly decrease the data rates of the NOMA transmission and even result in denial of service attacks.

Game theoretic study of anti-jamming communications in wireless networks such as [9]–[13] has provided insights into the defense against smart jammers. In this paper, the anti-jamming transmissions of a base station (BS) in an MIMO NOMA system are formulated as a zero-sum Stackelberg game. In this game, the transmit power of the BS for each user is first determined on each antenna and then a smart jammer chooses the jamming power in an attempt to interrupt the NOMA transmission at a low jamming cost. The Stackelberg equilibrium (SE) of the anti-jamming MIMO NOMA transmission game is derived, showing that the BS tends to allocate more power to the strong user but has to satisfy the minimum rate demand of weak users under full jamming-power attack of the smart jammer. Conditions under which the SE exists in the game are provided to disclose the impact of the radio channel states and the number of antennas on the communication efficiency of NOMA transmissions.

Reinforcement learning techniques, such as Q-learning can derive the optimal strategy with probability one, if all the feasible actions are repeatedly sampled over all the states in the Markov decision process [14]. Therefore, we propose a Q-learning based power allocation strategy that chooses the transmit power based on the observed state of the radio environment and the jamming power and a quality function or Q-function that describes the discount long-term reward for each state-action pair. This scheme is applied for the BS to derive the optimal policy for multiple users in the dynamic anti-jamming MIMO NOMA game without being aware of the channel model and the jamming model. A hotbooting Q-learning based power allocation exploits the experiences in similar anti-jamming NOMA transmission scenarios to set

Copyright (c) 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org. This work was supported in part by the National Natural Science Foundation of China under Grant 61671396 and in part by the U.S. National Science Foundation under Grants CCF-1420575, CNS-1456793, ECCS-1307949 and EARS-1444009.
the initial value of the quality function for each state-action pair and thus accelerates the convergence of the standard Q-learning algorithm that always uses zero as the initial Q-function value. Meanwhile, the Dyna architecture that formulates a learned world model from real experience can be used to increase the learning speed to derive the optimal policy [15]. The Dyna algorithm based power allocation scheme emulates the planning and reactions from hypothetical experiences and takes extra computation for updating Q-function values. Simulation results show that the proposed power allocation schemes enable the BS to improve the sum data rate for the users in the downlink NOMA transmission against jamming.

The contributions of this work can be summarized as follows:

- We formulate an anti-jamming MIMO NOMA transmission game and derive the SEs of the game.
- We propose a hotbooting Q-learning based power allocation scheme for multiple users in the dynamic NOMA transmission game against a smart jammer without knowing the jamming and channel models.
- The Dyna architecture and hotbooting techniques are applied to accelerate the learning speed and thus enhance the anti-jamming NOMA communication efficiency.

The remainder of the paper is organized as follows: The related work is reviewed in Section II. The anti-jamming NOMA transmission game is formulated in Section III, and the SE of the game is derived in Section IV. Two reinforcement learning based power allocation schemes are proposed for the dynamic NOMA transmission game in Section V and VI, respectively. Simulation results are provided in Section VII. The conclusions are drawn in Section VIII.

II. RELATED WORK

Power allocation is critical for NOMA transmissions. For instance, the low-complexity scheme as proposed in [16] improves the ergodic capacity for a given total transmit power constraint. The NOMA power allocation as proposed in [17] increases the sum data rates by pairing users with distinct channel conditions. The joint power allocation and user pairing method proposed in [4] employs the minorization-maximization algorithm to further improve the sum data rates in the multiple-input and single-output NOMA system. The power allocation method for multicarrier NOMA transmission designed in [18] exploits the heterogeneity of service quality demand to perform the successive interference cancellation and reduce power consumption.

User fairness is one of the key challenges in the NOMA power allocation. The power allocation based on proportional fairness proposed in [19] improves both the communication efficiency and user fairness in a two-user NOMA system. A bisection-based iterative NOMA power allocation developed in [20] enhances user fairness with partial channel state information. The cognitive radio inspired power allocation developed in [5] satisfies the quality-of-service requirements for all the users in the NOMA system. A NOMA system that pairs users with independent and identically distributed random channel states was investigated in [21] to reduce the outage probabilities for each user. However, most existing NOMA power allocation strategies do not provide jamming resistance, especially against smart jamming.

Game theoretic study of anti-jamming communications provides insights into the design of the power allocation [9]–[11], [22]–[26]. For example, the power allocation game formulated in [9] provides a closed-form solution to improve the ergodic capacity of MIMO under smart jamming. The MIMO transmission game formulated in [10] studies the power allocation strategy on the artificial noise against a smart attacker who can choose active eavesdropping or jamming. The anti-jamming transmission in a peer-to-peer network was formulated as a Bayesian game in [11], in which each jammer is identified based on the attack in the previous time slots. The stochastic communication game as analyzed in [25] proposes an intrusion detection system to assist the transmitter and increase the network capacity against both eavesdropping and jamming.

Learning based attacker type identification developed in [27] improves the transmission capacity under an uncertain type of attacks in the time-varying radio channels for the stochastic secure transmission game. The Q-learning based power control proposed in [28] resists smart jamming in cooperative cognitive radio networks. The hotbooting learning technique as proposed in [29] exploits the historical experiences to initialize the Q values and thus accelerates the convergence rate of Q-learning. Model learning based on tree structures as proposed in [30] further improves the sample efficiency in stochastic environments and thus the convergence rate of the learning process.

The Q-learning based power control strategy proposed in [31] improves the secure capacity of the MIMO transmission against smart attacks. Compared with our previous work in [31], we consider the anti-jamming MIMO NOMA transmission and improve the game model by incorporating the sum data rates and the minimum data rates required by each user in the utility function. In addition, a fast Q-based power allocation algorithm that combines the hotbooting technique and Dyna architecture is proposed to improve the communication efficiency compared with the standard Q-learning based scheme as developed in [31].

III. ANTI-JAMMING MIMO NOMA TRANSMISSION GAME

As shown in Fig. 1, we consider a time-slotted NOMA system consisting of $M$ mobile users each equipped with $N_T$ antennas to receive the signals from a BS against a smart jammer. The BS uses $N_T$ transmit antennas to send the signal denoted by $x^k_m$ to User $m$, with $m = 1, 2, \cdots, M$ and $k = 1, 2, 3, \cdots$. A superimposed $N_T$-dimensional signal vector $x^k = \sum_{n=1}^{M} x^k_n$ is sent at time $k$.

The smart jammer uses $N_J$ antennas to send jamming signal denoted by $z_j$ with the power constraint $p_j^k = \mathbb{E}[z_j^2]$. The jamming power $p_j^k \in [0, P_J]$ is chosen based on the ongoing downlink transmit power, where $P_J$ is the maximum jamming power.

Let $g^k_{j,i,m}$ denote the channel power gain from the $i$-th antenna of the BS to the $j$-th antenna of User $m$, and
The channel matrix $\mathbf{H}_{B,m}^{k}$ denoted by $[g_{j,i,m}^{k}]_{1 \leq j \leq N_{T}, 1 \leq i \leq N_{R}}$ represent the channel matrix between the BS and User $m$. The time index $k$ is omitted if no confusion occurs. Without loss of generality, the channel conditions of users are sorted as $||\mathbf{H}_{B,1,M}|| > \cdots > ||\mathbf{H}_{B,1,1}||$, while $|X|$ denotes the Frobenius norm of a matrix $X$, i.e., the user rank follows the order of channel power gains, indicating that User $M$ may stay in the cell central area while User 1 is at the cell-edge. More specifically, the MIMO channel can be viewed as a bundle of independent sub-channels and thus the channel gain matrices of the users can be ordered by the squared Frobenius norm, according to [4] and [32]. Similarly, $\mathbf{H}_{j,m}^{k}$ denotes the channel matrix between the jammer and User $m$. Each channel matrix is assumed to have independently and identically distributed (i.i.d.) complex Gaussian distributed elements, i.e., $\mathbf{H}_{\mu,m}^{k} \sim \mathcal{CN}(0, \sigma_{\mu,m}^{2})$, with $\mu = B, J$ and $m = 1, 2, \cdots, M$, and the $i$-th largest eigenvalue of $\mathbf{H}_{\mu,m}^{k}\mathbf{H}_{\mu,m}^{*}$ is denoted by $h_{\mu,i}^{m}$. Thus, the signal received by User $m$ is given by

$$y_{m} = \mathbf{H}_{B,m} \sum_{n=1}^{M} \mathbf{x}_{n} + \mathbf{H}_{J,m} \mathbf{z}_{J} + \mathbf{n}_{m}, \quad m = 1, 2, \cdots, M,$$

where the noise vector $\mathbf{n}_{m}$ consists of $N_{R}$ additive normalized zero-mean Gaussian noise for User $m$.

Let $\theta_{m}^{k}$ be the power allocation coefficient for User $m$, i.e., the allocated transmit power is $P_{T}^{k} \theta_{m}^{k}$. A BS has difficulty obtaining instantaneous channel state information accurately in time in the time-varying wireless environment and thus equally allocates the transmit power $P_{T} = \frac{P_{T}}{N_{T}}$ over the $N_{T}$ antennas, as mentioned in [16]. On the other hand, the BS can obtain the statistical results regarding channel states of the users via the feedback from the users and decide the signal decoding order accordingly. Thus the transmit covariance matrix of User $m$ denoted by $\mathbf{Q}_{m}$ is given by

$$\mathbf{Q}_{m} = \mathbb{E}(\mathbf{x}_{m} \mathbf{x}_{m}^{H}) = \frac{\theta_{m}^{k} P_{T}^{k}}{N_{T}} \mathbf{I}_{N_{T}},$$

where a superscript $H$ denotes the conjugate transpose. In this game, the power allocation vector chosen by the BS is denoted by $\theta^{k} = [\theta_{m}^{k}]_{1 \leq m \leq M}$ for all the $M$ users with the total power constraint $\sum_{1 \leq m \leq M} \theta_{m}^{k} = 1$. For simplicity, let $\Omega$ denote the vector space of all the available power allocation strategies, with $\forall \theta \in \Omega$.

On the other hand, without knowing the downlink channel power gain of the users, the jammer has to equally allocate the jamming power, with the jamming power covariance matrix given by

$$\mathbf{Q}_{J} = \mathbb{E}(\mathbf{z}_{J} \mathbf{z}_{J}^{H}) = \frac{P_{J}}{N_{J}} \mathbf{I}_{N_{J}}.$$  

The effective precoding and successive interference cancellation design for the MIMO-NOMA system have been studied in works such as [4] and [5]. More specifically, the downlink NOMA transmission scheme developed in [4] increases the sum data rate under the decodability constraint according to a simplified SIC method that orders the users based on the norm of the channel gain vector. The signal alignment scheme developed in [5] decomposes the multi-user MIMO-NOMA scenario into multiple separate single-antenna NOMA channels, and orders the users based on the large-scale channel fading gain with user pairing.

In this work, we apply the successive interference cancellation scheme that orders the users according to the channel power gains under jamming, similar to [16] and [33]. The decoding order according to the Frobenius norms of the channel matrices is optimal if the jamming power is not strong enough to change the SIC order, i.e., for any two users $j$ and $m$, $h_{m}^{B}/h_{m}^{J} \geq h_{j}^{B}/h_{j}^{J}$ holds if $m > j$. Nevertheless, if this assumption does not hold as the jamming power received by a user is much stronger than that received by the user who decodes afterwards, the NOMA system ranks the decoding order according to the channel power gain normalized by the jamming power in SIC, i.e., User $m$ decodes before User $j$, if $h_{m}^{B}/h_{m}^{J} \geq h_{j}^{B}/h_{j}^{J}$ with $m > j$, if not specified.

According to the singular value decomposition of the channel matrix $\mathbf{H}_{\mu,m}, \mu = B, J$, and the assumption of the Gaussian distributed signals, the achievable data rate of User $m$ denoted by $R_{m}$ depends on the SINR of the signal, and is given by


Thus the data rate of User \( M \) with the best channel condition can subtract the signals of all the other \( M-1 \) users from the superimposed signal \( x \). Thus the data rate of User \( M \) denoted by \( R_M \) is given by

\[
R_M = \log_2 \det \left( I + \left( I + H_{I,M} \mathbf{Q}_M \mathbf{H}_{B,M}^H \right)^{-1} H_{B,M} \mathbf{Q}_M \mathbf{H}_{B,M}^H \right).
\]

(4)

Applying interference cancellation as illustrated in [33], User \( M \) with the best channel condition can subtract the signals of all the other \( M-1 \) users from the superimposed signal \( x \). Thus the data rate of User \( M \) denoted by \( R_M \) is given by

\[
R_M = \log_2 \det \left( I + \left( I + H_{I,M} \mathbf{Q}_M \mathbf{H}_{B,M}^H \right)^{-1} H_{B,M} \mathbf{Q}_M \mathbf{H}_{B,M}^H \right).
\]

(5)

Note that the received jamming signal is viewed as noise, and does not change the decoding sequence according to the successive interference cancellation strategy in [33].

By exploiting the difference of the channel power gains between the users, the BS usually increases the power allocation coefficient for weak users to satisfy the quality of service (QoS) required by the user, such as the minimum data rate demand of the user denoted by \( R_0 \). Therefore, the NOMA transmission has to ensure \( \min(R_1, \ldots, R_M) \geq R_0 \).

The jammer aims to interrupt the NOMA transmission and make at least one user’s data rate less than the quality of service requirement, i.e., \( R_m < R_0 \), for \( 1 \leq m \leq M \), as shown in the first term in (6). If failing to do that, the jammer has a goal to reduce the overall data rate with less jamming power and avoid being detected, as indicated in (6). Let \( \gamma \) denote the jamming cost, including the jamming power consumption and the risk to be detected, which depends on the intrusion detection methods used by the BS.

The anti-jamming NOMA transmissions are formulated as a zero-sum Stackelberg game denoted by \( \mathcal{G} \), in which the BS as the leader first chooses the power allocation coefficient vector \( \theta = [\theta_m]_{1 \leq m \leq M} \) to improve the system throughput under the user rate constraint, and a smart jammer as the follower chooses the jamming power in an attempt to interrupt the NOMA transmission at a low jamming cost.

The utility of the BS denoted by \( u \) depends on the sum data rate of the \( M \) users and the jamming cost, and is given by

\[
u = 1 \left( \min_{1 \leq m \leq M} R_m \geq R_0 \right) \left( \sum_{m=1}^{M} R_m + \gamma p_j \right).
\]

(6)

where \( I(\cdot) \) is the indicator function that takes value 1 if the event is true and 0 otherwise, and the transmission fails if the QoS is not satisfied. In summary, we have considered an anti-jamming NOMA transmission game given by \( \mathcal{G} = \{ \{B, J\}, \{\theta, p_j\}, \{u, -u\} \} \). For ease of reference, we also summarize our commonly used notation in Table 1.

IV. SE OF THE NOMA TRANSMISSION GAME

At a Stackelberg equilibrium of the anti-jamming NOMA transmission game, the BS as the leader chooses its power allocation strategy for all downlink users first to maximize its utility given by (6) considering the response of the smart jammer as the follower. Then the jammer as the follower decides the jamming power to minimize the utility \( u \) based on the observed ongoing transmission. The SE strategies of the NOMA transmission game denoted by \( (\theta^*, p_j^*(\theta)) \) are given by definition as

\[
p_j^*(\theta) = \arg \min_{0 \leq P_j \leq P_f} u(\theta, p_j)
\]

(7)

\[
\theta^* = \arg \max_{\theta \in \Omega} u(\theta, p_j^*(\theta)).
\]

(8)

In the following, we first consider the transmission game for the BS with \( N_T \times N_R \) MIMO that serves \( M \) users, and then the specific case with 2 users to evaluate the SE strategies under different anti-jamming NOMA transmission scenarios.

Lemma 1. The anti-jamming \( N_T \times N_R \) NOMA transmission game with \( M \) users has a unique SE \( (\theta^*, p_j^*(\theta)) \) given by (9)-(10), if \( h_{m,i}^B h_{M,i}^J < h_{m,i}^B h_{m,i}^J, \forall 1 \leq m \leq M-1, \forall 1 \leq i \leq N_R \) and (11) holds.

Proof. See Appendix A.

According to (9)-(10), the SE of the anti-jamming NOMA transmission game depends on the QoS in the transmission \( R_0 \), the number of antennas \( (N_T \times N_R) \), the channel condition of the weak users and the maximum jamming power \( P_f \). If the jammer has good channel conditions to the weaker users, the BS allocates more power to the strongest user (i.e., User \( M \)) to improve the sum data rates, and meets the QoS of the weaker users. If the total transmit power is low for the transmission, the jammer applies full jamming power to block all the \( M \) users in terms of the transmission QoS.

Corollary 1. The anti-jamming \( N_T \times N_R \) NOMA transmission game \( \mathcal{G} \) with \( M \) users has a unique SE \( (\theta^*, 0) \), where \( \theta^* \) is given by \( \forall 1 \leq m \leq M-1 \)

\[
\sum_{i=1}^{N_R} \log_2 \left( 1 + \frac{\theta_m^* P_J h_{m,i}^B}{N_T + \left( 1 - \sum_{n=1}^{m} \theta_n^* P_J h_{m,i}^B \right)} \right) = R_0,
\]

(18)

if \( h_{m,i}^B h_{M,i}^J < h_{m,i}^B h_{m,i}^J, \forall 1 \leq m \leq M-1, \forall 1 \leq i \leq N_R \) and

\[
P_f \sum_{i=1}^{N_R} \left( N_T + \sum_{m=1}^{M-1} \theta_m^* P_J h_{m,i}^B \right) + \left( 1 - \sum_{n=1}^{M-1} \theta_n^* \right) P_J h_{M,i}^B < N_J \gamma \ln 2.
\]

(19)

Proof. See Appendix B.

If the jamming cost is high or the jamming channel experiences serious fading as shown in Eq. (19), a smart jammer...
will keep silent to reduce the power consumption. In addition, as the jamming power at SE also decreases with $N_T$, MIMO significantly improves the jamming resistance.

**Corollary 2.** The anti-jamming $N_T \times N_R$ NOMA transmission game $G$ with $M = 2$ users has a unique SE $(\theta_i^*, p_i^*)$ given by (12)-(13) if $h_{1,i}^j, h_{2,i}^j < h_{1,i}^B, h_{2,i}^B, \forall 1 \leq i \leq N_R$ and (14) holds.

**Proof.** See Appendix C.

As a concrete example, the utility of the BS in Fig. 2 is maximized if the rate demand of the weak user is satisfied under full jamming power (i.e., (12)). Moreover, the feasible allocation coefficient for the strong user increases with the total transmission power $P_T$ and the jamming power at the SE decreases with it.

Figure 3 presents the jamming power and the sum data rate of 2 users in terms of the channel power gain of User 1, $\sigma_{B,1}$. More specifically, the average SINR of users increases with $\sigma_{B,1}$ and the number of transmission antennas $N_T$, because the jammer will fail to interrupt the user experience and decrease the jamming power to reduce the cost. Therefore, the sum data rate of the $M$ users improves with $\sigma_{B,1}$ and $N_T$.

**Corollary 3.** The anti-jamming $N_T \times N_R$ NOMA transmission game $G$ with $M = 2$ users has a unique SE $(\theta_i^*, p_i^*)$ given by (15)-(16) if $h_{1,i}^j, h_{2,i}^j > h_{1,i}^B, h_{2,i}^B, \forall 1 \leq i \leq N_R$ and (17) holds.

**Proof.** See Appendix D.

If the jamming channel condition to the stronger user (i.e., User 2) is high as shown in the given condition, the sum data rate increases with more power allocated for the weak user (i.e., User 1). On the other hand, the smart jammer, as a follower observing the ongoing transmission will apply its full
jamming power to interrupt the communication of the stronger user if its transmit power is not high enough as in (15).

V. NOMA POWER ALLOCATION BASED ON HOTBOOTING Q-LEARNING

The repeat interactions between a BS and a smart jammer can be formulated as a dynamic anti-jamming NOMA transmission game. The optimal downlink power allocation depends on the locations of the users, the radio channel states and the jamming parameters in the time slot, which are challenging for a BS to accurately estimate. In addition, the power allocation strategy of the BS has an impact on the jamming strategy and the received SINR of the users and thus the anti-jamming NOMA transmission process can be formulated as a finite Markov decision process. Therefore, the NOMA system can apply Q-learning, a widely-used model-free reinforcement learning technique in the power allocation. However, the training data initialized usually requires exponentially more data than the optimal policy. Therefore, we propose a hotbooting technique, which initializes the Q-value based on the training data obtained in advance from large-scale experiments in similar scenarios. As we will see, the hotbooting Q-learning based power allocation can decrease the useless random explorations that reflects the environment dynamics, denoted by $s^k$, is chosen as the last received SINR of each user denoted by $\text{SINR}_{\text{m}}^{k-1}$, i.e., $s^k = \{\text{SINR}_{\text{m}}^{k-1}\}_{1 \leq \text{m} \leq M} \in \xi$, where $\xi$ is the space of all the possible SINR vectors. More specifically, each user quantizes the received SINR into one of $L$ levels for simplicity, and sends back the quantized value to the BS as the observed state for the next time slot. Therefore, the BS uses the $ML$ vector as feedback information to update its Q function in the Q-learning based NOMA system.

For simplicity, the feasible action set of the BS in the power allocation, or each of the transmit power coefficient is quantized into $L$-non-zero levels denoted by $\theta_{\text{m}}^{k}$, i.e., $\theta^k = \{\theta_{\text{m}}^{k}\}_{1 \leq \text{m} \leq M} \in \Omega$. The transmit signal of the $M$ users is sent at the power allocation given by $\theta^k$.

Note that the random exploration of standard Q-learning at the beginning of the game due to the all-zero Q-value initialization usually requires exponentially more data than the optimal policy. Therefore, we propose a hotbooting technique, which initializes the Q-value based on the training data obtained in advance from large-scale experiments in similar scenarios. As we will see, the hotbooting Q-learning based power allocation can decrease the useless random explorations.
Algorithm 1: Hotbooting preparation

1: \[ Q(s, \theta) = 0, V(s) = 0, \forall s \in \xi, \forall \theta \in \Omega, \text{ and } s^0 = 0 \]
2: \[ \text{for } i = 1, 2, \ldots, I \text{ do} \]
3: \[ \text{Emulate a similar environment} \]
4: \[ \text{for } k = 1, 2, \ldots, K \text{ do} \]
5: \[ \text{Choose action } \theta^k \text{ at random} \]
6: \[ \text{Obtain the utility } u^k \text{ and the received SINR of each user } \text{SINR}_{1 \leq m \leq M}^k \]
7: \[ s^{k+1} = [\text{SINR}_{1 \leq m \leq M}^k] \]
8: \[ Q^*(s^k, \theta^k) = (1-\alpha)Q^*(s^k, \theta^k) + \alpha(u^k + \delta V^*(s^{k+1})) \]
9: \[ V^*(s^k) = \max_{\theta \in \Omega} Q^*(s^k, \theta) \]
10: \[ \text{end for} \]
11: \[ \text{end for} \]

Algorithm 2: Hotbooting Q-learning based NOMA power allocation

1: \[ \text{Set } Q(s, \theta) = Q^*(s, \theta), \forall s \in \xi, \forall \theta \in \Omega, \text{ and } s^0 = 0 \]
2: \[ \text{for } k = 1, 2, 3, \ldots \text{ do} \]
3: \[ \text{Choose } \theta^k \text{ via (22)} \]
4: \[ \text{for } m = 1, 2, \ldots, M \text{ do} \]
5: \[ \text{Allocate power } \theta^k_P \text{ for the signal to User } m \]
6: \[ \text{end for} \]
7: \[ \text{Send the superimposed signal } x^k \text{ over } N_T \text{ antennas} \]
8: \[ \text{Observe } \text{SINR}_{1 \leq m \leq M}^k \text{ and the utility } u^k \]
9: \[ s^{k+1} = [\text{SINR}_{1 \leq m \leq M}^k] \]
10: \[ \text{Update } Q(s^k, \theta^k) \text{ via (20)} \]
11: \[ \text{Update } V(s^k) \text{ via (21)} \]
12: \[ \text{end for} \]

The BS with our proposed RL based power allocation scheme does not know the instantaneous channel state information, the jamming power and the jamming channel gains, which are required by the global optimization algorithm such as [35]. In this way, the BS learns the jamming strategy according to the anti-jamming NOMA transmission history and derives the optimal transmit power allocation strategy that improves the long-term anti-jamming communication efficiency.

However, the convergence complexity of Algorithm 2 depends on the size of the action-state space \(|\xi \times \Omega|\), which exponentially increases with the number of users and the quantization levels of the power allocation coefficients. The channel time variation on the other hand, requires much faster learning rate of the power allocation. Therefore, we further propose a fast Q-learning based power allocation scheme that combines the hotbooting technique and the Dyna architecture to improve the communication efficiency of Algorithm 2 in dynamic radio environments in the following section.

VI. NOMA POWER ALLOCATION BASED ON FAST Q-LEARNING

The Dyna architecture formulates a learned world model from the real anti-jamming NOMA transmission experiences to accelerate the learning speed of Q-learning in dynamic radio environments. More specifically, we apply the Dyna architecture that emulates the planning and reactions from hypothetical experiences to improve the anti-jamming efficiency in the fast Q-learning based NOMA power allocation algorithm. The algorithm is summarized in Algorithm 3, and the main structure is illustrated in Fig. 4.

The fast Q-learning power allocation algorithm also applies Algorithm 1 to initialize the Q values with the hotbooting technique and updates the Q functions via (20) and (21) similarly to Algorithm 2 according to the NOMA transmission result in each time slot. The BS observes the received SINR of each user at last time slot denoted by SINR_{1 \leq m \leq M}^{k-1} as the state s^k, i.e., s^k = [\text{SINR}_{1 \leq m \leq M}^{k-1}] as the state \text{SINR}_{1 \leq m \leq M}^k. The power allocation strategy \theta^k is also chosen according to the system state and the Q-function based on the \varepsilon-greedy policy as shown in (22).
Meanwhile, the Dyna-Q based algorithm also uses the real experiences of the NOMA transmission at time $k$ to build a hypothetical anti-jamming NOMA transmission model. More specifically, the real experience of the NOMA transmission at time $k$ is saved as an experience record for the corresponding state-action pair, i.e., $(s^k, \theta^k)$. The experience record stored at the BS consists of the occurrence counter vector of the state-action pairs denoted by $\Phi$, the occurrence counter vector of the next state denoted by $\Phi'$, the reward record $\tau'$, and the modeled reward $\tau$. The hypothetical experience is generated according to the real experience $(\Phi', \Phi, \tau', \tau, \Psi)$. As shown in Algorithm 3, the counter vector of the next state $\Phi'$ is updated according to the anti-jamming NOMA transmission at time $k$, i.e.,

$$
\Phi'(s^k, \theta^k, s^{k+1}) = \Phi'(s^k, \theta^k, s^{k+1}) + 1.
$$

The occurrence counter vector in the hypothetical experience $\Phi$ defined as the sum of $\Phi'$ over all the feasible next state is then updated with

$$
\Phi(s^k, \theta^k) = \sum_{s^{k+1} \in \xi} \Phi'(s^k, \theta^k, s^{k+1}).
$$

The reward record from the real experience $\tau'$ in this time slot is the utility of the BS at time $k$ given by

$$
\tau'(s^k, \theta^k, \Phi(s^k, \theta^k)) = u(s^k, \theta^k).
$$

Based on the reward record $\tau'$ in (25) and the state transition from $s^k$ to $s^{k+1}$, we can formulate a hypothetical anti-jamming NOMA transmission model to generate several hypothetical NOMA transmission experiences, in which the reward to the BS denoted by $\tau$ is defined as the utility of the BS averaged over all the previous real experiences and is given by

$$
\tau(s^k, \theta^k) = \frac{1}{\Phi(s^k, \theta^k)} \sum_{\tau'} \tau'(s^k, \theta^k, n).
$$

The transition probability from $s^k$ to $s^{k+1}$ in the hypothetical experience denoted by $\Psi$ is given by (23) and (24) as

$$
\Psi(s^k, \theta^k, s^{k+1}) = \frac{\Phi'(s^k, \theta^k, s^{k+1})}{\Phi(s^k, \theta^k)}.
$$

The Q function is updated with $D$ more times with the Dyna architecture, in which $D$ hypothetical experiences are generated according to the world model $\Psi$ learned from the previous real experience $(\Phi', \Phi, \tau', \tau, \Psi)$. More specifically, the system state denoted by $s^d$ and the hypothetical action denoted by $\hat{\theta}^d$ in the $d$-th additional update are generated based on the transition probability $\Psi(s^d, \hat{\theta}^d, s^{d+1})$ given in (27), i.e.,

$$
\Pr(s^{d+1} | s^d, \hat{\theta}^d) = \Psi(s^d, \hat{\theta}^d, s^{d+1}),
$$

where $(s^0, \hat{\theta}^0)$ is randomly chosen from $\xi$ and $\Omega$, respectively.

The Q function is updated in the $d$-th hypothetical experience with $d = 1, 2, \ldots, D$ by the following:

$$
Q(s^d, \hat{\theta}^d) \leftarrow (1 - \alpha)Q(s^d, \hat{\theta}^d) + \alpha(\tau(s^d, \hat{\theta}^d) + \delta V(s^{d+1}))
$$

$$
V(s^d) = \max_{\theta \in \Omega} Q(s^d, \theta).
$$

As shown in Algorithm 3, the proposed NOMA power allocation algorithm requires an additional memory in each time slot to store the experience record $(\Phi', \Phi, \tau', \tau, \Psi)$. Compared with the standard Q-learning based power allocation in Algorithm 2, this scheme has $D$ more updates of the Q functions. However, the proposed fast Q-based power

---

**Algorithm 3: NOMA power allocation with fast Q-learning**

1. Set $Q(s, \theta) = Q(s, \theta)$, $\forall s \in \xi$, $\forall \theta \in \Omega$, and $s^0 = 0$
2. for $k = 1, 2, 3, \ldots$
3. Choose $\theta^k$ via (22)
4. for $m = 1, 2, \ldots, M$
5. Allocate power $\theta_m$ for the signal to User $m$
6. end for
7. $\theta^k = \left[ \theta_m^k \right]_{1 \leq m \leq M}$
8. Send $x^k$ according to $\theta^k$ over $N_T$ antennas
9. Observe SINR $s^k \leq s \leq M$ and the utility $u^k$
10. $s^{k+1} = \left[ \text{SINR}_m^k \right]_{1 \leq m \leq M}$
11. Update $Q(s^k, \theta^k)$ via (20)
12. Update $V(s^k)$ via (21)
13. Calculate $\Phi'(s^k, \theta^k, s^{k+1})$ via (23)
14. Calculate $\Phi(s^k, \theta^k)$ via (24)
15. Calculate $\Psi(s^k, \theta^k, s^{k+1})$ via (27)
16. Calculate $\tau(s^k, \theta^k)$ via (25)
17. Randomly select $s^d \in \xi$ and $\hat{\theta}^0 \in \Omega$
18. for $d = 1$ to $D$
19. $s^{d+1}$ according to (28)
20. Obtain $\tau(s^d, \hat{\theta}^d)$ via (26)
21. Update $Q(s^d, \hat{\theta}^d)$ via (29)
22. Update $V(s^d)$ via (30)
23. end for
24. end for
allocation algorithm applies the Dyna architecture to increase the convergence speed of the reinforcement learning process and thus improve the NOMA communication efficiency in the presence of smart jamming.

VII. Simulation Results

In this section, we evaluate the performance of the MIMO NOMA power allocation scheme for $M = 3$ users in the dynamic anti-jamming communication game via simulations. If not specified otherwise, we set $N_T = N_R = 5$, $\sigma_{B,1} = 10$ dB, $\sigma_{B,2} = 11$ dB, $\sigma_{B,3} = 20$ dB, $P_T = 20$ W, $R_0 = 1$ bps/Hz, $\gamma = 0.5$, $\alpha = 0.2$, $\delta = 0.7$, $\varepsilon = 0.9$ and $I = 200$. In the simulations, a smart jammer applies Q-learning to determine the jamming power according to the current downlink transmit power with $N_J = 3$, $\sigma_{J,1} = \sigma_{J,2} = 11$ dB, $\sigma_{J,3} = 12$ dB, and $P_J = 20$ W.

As shown in Fig. 5, the Q-learning based power allocation scheme has significantly increased the jamming resistance of NOMA transmissions, which can be further improved by the hotbooting technique and the Dyna architecture. In the benchmark Q-learning based OMA system, the BS equally allocates the frequency and time resources to each user and chooses the transmit power according to the current state and the Q function, similarly to Algorithm 2.

According to [36], the NOMA transmission outperforms OMA with less outage probability, which is verified by the simulation results in Fig. 5. More specifically, the Q-learning based NOMA system can improve the communication efficiency against jamming attack compared to the OMA system. For instance, the Q-learning based anti-jamming NOMA system exceeds the Q-learning based OMA system with 8.7% higher SINR, 42.8% higher sum data rate, and 7.1% higher utility of the BS at the 1000-th time slot.

The NOMA transmission performance can be further improved by the hotbooting technique. For instance, the NOMA system with hotbooting Q-learning achieves 12% higher SINR, 10% higher sum data rate, and 9% higher utility, at the 1000-th time slot. The reason is that the hotbooting technique can formulate the emulated experiences in dynamic radio environments to effectively reduce the exploration trials and thus significantly improve the convergence speed of Q-learning and the jamming resistance efficiency. The power allocation performance can be further improve by the Dyna architecture with hypothetical experiences. For instance, the average SINR, the sum data rate, and the utility of the proposed fast Q-learning algorithm are 11%, 10%, and 7% higher at the 1000-th time slot, compared with the hotbooting Q-learning scheme.

The performance of the proposed power allocation scheme in the $5 \times 5$ MIMO NOMA system with 2 users is evaluated with the varying channel gains of User 1 in Fig. 6. Both the sum capacity of users and the utility of the BS increase with the channel gain $\sigma_{B,1}$. For instance, as the channel gain $\sigma_{B,1}$ increases from 4 to 12 dB, the average SINR, the sum data rate and the utility of the Q-learning based power allocation increase by 9%, 14% and 17%, respectively. If $\sigma_{B,1} = 12$ dB, the fast Q-based power allocation achieves the best anti-jamming communication and performance exceeds
Fig. 6. Performance of the power control schemes in the dynamic MIMO NOMA transmission game in the presence of smart jamming versus the channel gain of User 1, with $M = 2$, $N_T = N_R = 5$, $N_J = 3$, $\sigma_{J,1} = \sigma_{J,2} = 12$ dB, $P_T = P_J = 20$ W, $R_0 = 2$ bps/Hz and $\gamma = 0.5$. 

Fig. 7. Performance of the power control schemes in the dynamic MIMO NOMA transmission game in the presence of smart jamming versus the number of the RX antennas, with $M = 2$, $N_T = 10$, $\sigma_{B,1} = 12$ dB, $\sigma_{B,2} = 20$ dB, $\sigma_{J,1} = \sigma_{J,2} = 11$ dB, $P_T = P_J = 20$ W, $R_0 = 2$ bps/Hz and $\gamma = 0.5$. 

(a) Average SINR

(b) Sum data rates

(c) Utility of the BS

(a) Average SINR

(b) Sum data rates

(c) Utility of the BS
the standard Q-learning based strategy with 9% higher SINR, 14% higher sum capacity, and 22% higher utility.

Fig. 7 shows that the NOMA transmission efficiency improves with the number of the receive antennas with $N_R = 10$ transmit antennas. The proposed power allocation schemes have strong resistance against smart jamming even with a large number of jamming antennas $N_J$. For instance, our proposed scheme can improve the average SINR, the sum data rate and the user utility of the $10 \times 8$ NOMA system against a jammer with 2 antennas by 10%, 12%, and 15%, respectively, compared with the benchmark strategy.

VIII. CONCLUSION

In this paper, we have formulated an anti-jamming MIMO NOMA transmission game, in which the BS determines the transmit power to improve its utility based on the sum data rate of the users, and the smart jammer as the follower in the Stackelberg game chooses the jamming power at an attempt to interrupt the ongoing transmission at a low jamming cost. The SE of the game has been derived, and the conditions assuring its existence have been provided, showing how the anti-jamming communication efficiency of NOMA systems increases with the number of antennas and the channel power gains. A fast Q-based NOMA power allocation scheme that combines the hotbooting technique and Dyna architecture is proposed for a dynamic game to accelerate the learning and thus improve the communication efficiency against smart jamming. As shown in the simulation results, the proposed NOMA power allocation scheme can significantly improve the average SINR, the sum data rate and the utility of the BS soon after the start of the game. For example, the sum data rate of 3 users in the $5 \times 5$ MIMO NOMA anti-jamming transmission game increases by 128% after 1000 time slots, which is 21% higher than that of the standard Q-learning based scheme.

We have analyzed a simplified jamming scenario for MIMO NOMA systems in this work. A future direction of our work is to extend the theoretical analysis of the NOMA transmission to more practical scenarios with smart jamming, in which a jammer uses programmable radio devices to flexibly choose multiple jamming policies.

REFERENCES


u = \sum_{i=1}^{N_R} \left( \sum_{m=1}^{M-1} \log_2 \left( 1 + \frac{N_J \theta_m P_T h_{B,i}^m}{N_T N_J + N_J \left( 1 - \sum_{n=1}^{M-1} \theta_n \right) P_T h_{B,i}^m + N_T \bar{p} h_{l,i}^m} \right) \right) + \log_2 \left( 1 + \frac{N_J \left( 1 - \sum_{n=1}^{M-1} \theta_n \right) P_T h_{B,i}^m}{N_T N_J + N_T \bar{p} h_{l,i}^m} \right) + \gamma p_J, \tag{31}

\frac{\partial u(\theta, p_J)}{\partial \theta_m} = \sum_{i=1}^{N_R} \frac{N_J P_T \left( N_T p_J^* \left( h_{B,i}^m h_{M,i}^J - h_{B,i}^m h_{l,i}^J \right) - N_J P_T h_{B,i}^m h_{M,i}^J \sum_{n=1}^{M-1} \theta_n - N_J N_J \left( h_{B,i}^m - h_{l,i}^m \right) \right) \right) \left( N_T N_J + N_T p_J^* h_{l,i}^J + N_J \left( 1 - \sum_{n=1}^{M-1} \theta_n \right) P_T h_{B,i}^m \right)}{\ln 2 \left( N_T N_J + N_T p_J^* h_{l,i}^J + N_J \left( 1 - \sum_{n=1}^{M-1} \theta_n \right) P_T h_{B,i}^m \right)} \left( N_T N_J + N_T p_J^* h_{l,i}^J + N_J \left( 1 - \sum_{n=1}^{M-1} \theta_n \right) P_T h_{B,i}^m \right) < 0, \forall 1 \leq m \leq M - 1. \tag{32}

u(\theta, p_J) = 0. Therefore, u is maximized if \( R_m(\theta, P_J) = R_0 \), \( \forall 1 \leq m \leq M - 1 \), i.e. that \( \theta^* \) is given by (9) and thus (8) holds. Therefore, the SE (\( \theta^*, p_J^* \)) is given by (9)-(10).

APPENDIX B
PROOF OF COROLLARY 1

Similar to the proof of Lemma 1, if \( \min \left[ R_m \right]_{1 \leq m \leq M} \geq R_0 \) for \( \forall 0 \leq p_J \leq P_J \), by (31), \( \partial^2 u(\theta, p_J)/\partial p_J^2 > 0 \). By (19), we have \( \partial u(\theta, p_J)/\partial p_J |_{p_J = 0} > 0 \). Thus \( u(\theta, p_J) \) is minimized at \( p_J = 0 \), and we have \( p_J^* = 0 \) by Eq. (7). If \( h_{B,i}^m h_{M,i}^J < h_{B,i}^m h_{l,i}^J, \forall 1 \leq m \leq M - 1, \forall 1 \leq i \leq N_R \), we have (32) and thus \( \theta^* \) given by (18). Therefore, if \( h_{B,i}^m h_{M,i}^J < h_{B,i}^m h_{l,i}^J, \forall 1 \leq m \leq M - 1, \forall 1 \leq i \leq N_R \) and (19) holds, we have the SE (\( \theta^*, 0 \)).

APPENDIX C
PROOF OF COROLLARY 2

By (33), if \( h_{B,i}^m h_{l,i}^J < h_{l,i}^J h_{B,i}^J, \forall 1 \leq i \leq N_R \), and (14) holds, we have \( \partial u(\theta_1, p_J^*)/\partial \theta_1 < 0 \), i.e. (34). Thus we have that \( u \) monotonically increases with \( \theta_2, 0 \leq \theta_2 \leq 1 \). On the other hand, it is clear that \( R_1(\theta_1, p_J^*) = 0 \) if \( \theta_1 = 0 \), and \( R_1(\theta_1, p_J^*) \) monotonically increases with \( \theta_1 \). Thus \( u \) is maximized if \( R_1 = R_0 \), i.e. that \( \theta_1^* \) is given by (12), and Eq. (8) is satisfied with \( \theta_1^* \). Therefore, the SE (\( \theta_1^*, p_J^* \)) is given by (12)-(13).

APPENDIX D
PROOF OF COROLLARY 3

By (32), if \( h_{B,i}^m h_{l,i}^J > h_{l,i}^J h_{B,i}^J + (\theta_1^* - \theta_1^*) N_J/p_J^*, \forall 1 \leq i \leq N_R \), and (17) holds, we have \( \partial u(\theta_1, p_J^*)/\partial \theta_1 < 0 \). It is clear that \( R_2(\theta_1, p_J^*) = 0 \) if \( \theta_1 = 1 \), and \( R_2(\theta_1, p_J^*) \) monotonically decreases with \( \theta_1 \). Thus \( u \) is maximized if \( R_2 = R_0 \), i.e. that \( \theta_1^* \) is given by (15). Therefore, the SE (\( \theta_1^*, p_J^* \)) is given by (15)-(16).
Liang Xiao (M’09, SM’13) is currently a Professor in the Department of Communication Engineering, Xiamen University, Fujian, China. She has served as an associate editor of IEEE Trans. Information Forensics and Security and guest editor of IEEE Journal of Selected Topics in Signal Processing. She is the recipient of the best paper award for 2016 INFOCOM Big Security WS and 2017 ICC. She received the B.S. degree in communication engineering from Nanjing University of Posts and Telecommunications, China, in 2000, the M.S. degree in electrical engineering from Rutgers University, NJ, in 2003, and the Ph.D. degree in electrical engineering from Tsinghua University, China, in 2009. She was a visiting professor with Princeton University, Virginia Tech, and University of Maryland, College Park.

Yanda Li received the B.S. degree in communication engineering from Xiamen University, Xiamen, China, in 2013, where he is currently pursuing the M.S. degree with the Department of Communication Engineering. His research interests include network security and wireless communications.

Canhuang Dai received the B.S. degree in communication engineering from Xiamen University, Xiamen, China, in 2017, where he is currently pursuing the M.S. degree with the Department of Communication Engineering. His research interests include smart grid and wireless communications.

Huaiyu Dai (F’17) received the B.E. and M.S. degrees in electrical engineering from Tsinghua University, Beijing, China, in 1996 and 1998, respectively, and the Ph.D. degree in electrical engineering from Princeton University, Princeton, NJ in 2002. He was with Bell Labs, Lucent Technologies, Holmdel, NJ, in summer 2000, and with AT&T Labs-Research, Middletown, NJ, in summer 2001. He is currently a Professor of Electrical and Computer Engineering with NC State University, Raleigh. His research interests are in the general areas of communication systems and networks, advanced signal processing for digital communications, and communication theory and information theory. His current research focuses on networked information processing and cross-layer design in wireless networks, cognitive radio networks, network security, and associated information-theoretic and computation-theoretic analysis. He has served as an editor of IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON SIGNAL PROCESSING, and IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. Currently he is an Area Editor in charge of wireless communications for IEEE TRANSACTIONS ON COMMUNICATIONS. He co-edited two special issues of EURASIP journals on distributed signal processing techniques for wireless sensor networks, and on multiuser information theory and related applications, respectively. He co-chaired the Signal Processing for Communications Symposium of IEEE Globecom 2013, the Communications Theory Symposium of IEEE ICC 2014, and the Wireless Communications Symposium of IEEE Globecom 2014.

H. Vincent Poor (S’72, M’77, SM’82, F’87) received the Ph.D. degree in electrical engineering and computer science from Princeton University in 1977. From 1977 until 1990, he was on the faculty of the University of Illinois at Urbana-Champaign. Since 1990 he has been on the faculty at Princeton, where he is the Michael Henry Strater University Professor of Electrical Engineering. During 2006 to 2016, he served as Dean of Princeton’s School of Engineering and Applied Science. He has also held visiting appointments at several other institutions, most recently at Berkeley and Cambridge. His research interests are in the areas of information theory and signal processing, and their applications in wireless networks, energy systems and related fields. Among his publications in these areas is the recent book Information Theoretic Security and Privacy of Information Systems (Cambridge University Press, 2017). Dr. Poor is a member of the National Academy of Engineering and the National Academy of Sciences, and is a foreign member of the Chinese Academy of Sciences, the Royal Society and other national and international academies. Recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal, Honorary Professorships from Peking University and Tsinghua University, both conferred in 2017, and a D.Sc. honoris causa from Syracuse University awarded in 2017.