Multiple Correlated Jammers Suppression: A Deep Dueling Q-Learning Approach

Linh Manh Hoang, Diep N. Nguyen, J. Andrew Zhang, and Dinh Thai Hoang
School of Electrical and Data Engineering, University of Technology Sydney, Australia

***Abstract*—For wireless networks under jamming attacks, suppressing the jammer is essential to guarantee a reliable communication link. However, it can be problematic to nullify the jamming signal when the correlations between transmitted jamming signals are deliberately varied over time. Specifically, recent studies reveal that the time-varying correlations create a "virtual change" in the jamming channel and thus their nullspace, even when the physical channels remain unchanged. Unlike existing studies that only consider unchanged correlations or merely propose a heuristic adaptation to the changing correlation problem by continuously monitoring the residual jamming signal then updating the beam-forming matrix, we develop a deep dueling Q-learning framework to minimize the magnitude of the "virtual change" by tuning the duration for different phases of each communication frame. Extensive simulations show that the proposed techniques can suppress the jamming signal, even when the correlations vary over time, and the correlations' range is unknown. It is worth noting that techniques do not require frequent monitoring of the residual jamming signals (after the nullification process) before updating the beam-forming matrix. As such, the system is more spectral-efficient and has a reduced outage probability.**

***Index Terms*—Correlated jamming, jamming suppression/nullification, deep dueling, Q-learning, frame adaptation.**

I. Introduction

The angle of arrival (AOA)-based beam-forming technique is a conventional approach to suppress jamming signals. It is accomplished by first estimating the AOAs of spatial streams of jamming signals, and then forming receiving beam nulls towards the estimated AOAs. However, at least one degree-of-freedom is needed to nullify each propagation path of the jamming signal [1]. Therefore, this method is only applicable when the number of jammers is small, and the environment is sparse scattering.

Another method to suppress the jamming signal is by estimating jamming channels characteristics, such as their nullspace [2], their projection [3], their ratios [4], and then deriving filters to suppress the jamming signals. These techniques require only a single degree-of-freedom to suppress each jammer, thus are more efficient than the aforementioned AOA-based approach. However, [2]–[4] do not evaluate the impacts of the time-varying correlations between transmitted jamming signals on the performance of their techniques. In [5] and [6], it is shown that the jammers can dramatically escalate the jamming impact by precisely choosing the correlations. In [7], the authors prove that the time-varying correlations create a "virtual change" in the jamming channel and hence their nullspace, making the beam-forming matrix derived from the estimated nullspace of the jamming channel incapable of nullifying jamming signals. [7] also proposes a heuristic solution to continuously monitor the residual jamming signals (after applying the nullification technique) and then adjust the beam-forming matrix when the residual surpasses a predefined value. However, the jamming residual monitoring process incurs additional overhead to the system, thus significantly reducing the spectral efficiency.

This paper aims to improve the system's spectral efficiency by minimizing the amount of time spent updating the beam-forming matrix, especially when the jammers use time-varying correlations between transmitted jamming signals, and the correlation values' range is unknown at the BS and the UEs. The jammers can even deliberately change the correlations range, making jamming nullification even more challenging. To deal with such uncertainty and incomplete information, we design a deep dueling Q-learning algorithm to minimize the magnitude of the "virtual change", thereby ensuring the effectiveness of the beam-forming matrix against this change. To the best of our knowledge, this is the first study to resolve the "virtual change" problem without constantly monitoring the residual jamming signal and then updating the beam-forming matrix. Our technique costs only a single degree-of-freedom to nullify each jammer, while remaining capable of nullifying jamming signals, even with an unknown and varying jamming strategy. Simulation results show that our technique achieves significantly higher system's spectral efficiency and a lower outage probability.

Notations: We use $(.)^H$ for Hermitian matrix transpose, and $|.|$ for complex number's modulus, respectively.

II. System Model

A. Network Model

We consider a multi-user multiple-input multiple-output (MU-MIMO) downlink system with one BS and K user equipment (UE). The number of antennas at the kth UE and the BS are N_k and N_T, respectively. The BS-UEs communication system is interfered by N_J single-antenna proactive jammers. Note that because a multi-antenna jammer can be treated as multiple single-antenna jammers, our technique is readily extendable to the case with multi-antenna jammer.

B. Signal Model

The received signals at the kth UE can be written as

$$\mathbf{y}_k = \mathbf{H}_k \mathbf{x}_k + \mathbf{H}_k \sum_{i \neq k}^{K} \mathbf{x}_i + \mathbf{Z}_k \mathbf{x}_{\mathrm{J}} + \mathbf{n}, \tag{1}$$

where $\mathbf{H}_k$ is the BS-kth UE channel, $\mathbf{x}_k$ is the transmitted signal targeted to the kth UE, $\mathbf{Z}_k$ denotes the channel from the N_J jammers to the kth UE, $\mathbf{x}_{\mathrm{J}}$ denotes the transmitted jamming signals, and $\mathbf{n}$ is the complex noise. We assume $\mathbf{n} \sim$

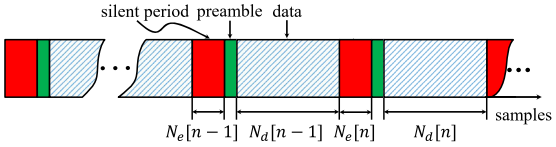

Fig. 1: Frame structure.

$\mathcal{CN}(\mathbf{0}, \sigma_n^2 \mathbf{I}_{N_k})$, where $\mathbf{I}_{N_k}$ is the identity matrix of size N_k, and σ_n^2 is the noise variance. Likewise, $\mathbf{x}_k \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Sigma}_{\mathbf{x}_k})$, where $\mathbf{\Sigma}_{\mathbf{x}_k}$ is the covariance matrices of $\mathbf{x}_k$. We assume $\mathbf{x}_\mathrm{J} \sim \mathcal{CG}(\boldsymbol{\mu}_\mathrm{J}, \mathbf{\Sigma}_\mathrm{J})$ [7], where $\mathcal{G}$ is a complex distribution function concealed from the UEs and the BS, $\boldsymbol{\mu}_\mathrm{J}$ and $\mathbf{\Sigma}_\mathrm{J}$ are the transmitted jamming signals' mean and covariance matrix, respectively.

III. Problem Formulation

A. Communication Protocol

To sustain the communication between the BS and kth UE under jamming, one can either adapt to the jammer (e.g., using rate adaptation or frequency hopping [8]) or suppress the jamming signals. In this paper, we focus on nullifying the jamming signals. Fig. 1 illustrates the communication protocol to deal with the jamming signals. The whole communication process is divided into consecutive frames, each contains three phases: estimation, preamble, and data transmission phases.

- During the estimation phase, which lasts for N_e samples, the BS does not transmit any signal, and the beam-forming matrix $\mathbf{W}_k$ is estimated. The beam-forming matrix $\mathbf{W}_k$ is utilized to suppress the jamming signal, and is designed by choosing B_k rows of $\mathbf{W}_k$ from $(N_k - N_J)$ rows of $\hat{\mathbf{G}}_k$. Note that $\hat{\mathbf{G}}_k$ denotes the estimated value of $\mathbf{G}_k$, and $\mathbf{G}_k$ is a matrix whose rows form an orthonormal basis for the left nullspace [9] of the received jamming signals. To exploit all of the remaining degree-of freedom (after jamming suppression) for BS-UEs communication, we use all the rows of $\hat{\mathbf{G}}_k$, letting $\mathbf{W}_k = \hat{\mathbf{G}}_k$. The algorithm to estimate $\mathbf{G}_k$ is presented below.
  Let $\mathbf{y}^\mathrm{e}_{\mathrm{J_k}}$ be the received jamming signal at the kth UE during the estimation phase when $\hat{\mathbf{G}}_k$ is estimated. Let $\mathbf{R}^\mathrm{e}_{\mathrm{J_k}}$ be a sample covariance matrix of $\mathbf{y}^\mathrm{e}_{\mathrm{J_k}}$, we have,
  $$\mathbf{y}^\mathrm{e}_{\mathrm{J_k}} = \mathbf{Z}_k \mathbf{x}^\mathrm{e}_\mathrm{J} + \mathbf{n}, \tag{2}$$
  $$\mathbf{R}^\mathrm{e}_{\mathrm{J_k}} = \frac{1}{N_\mathrm{e}} \mathbf{Y}^\mathrm{e}_{\mathrm{J_k}} (\mathbf{Y}^\mathrm{e}_{\mathrm{J_k}})^H, \tag{3}$$
  where $\mathbf{Y}^\mathrm{e}_{\mathrm{J_k}}$ is a set composed of N_e samples of $\mathbf{y}^\mathrm{e}_{\mathrm{J_k}}$. We obtain $\hat{\mathbf{G}}_k$ by $\hat{\mathbf{G}}_k = (\mathbf{U}_\mathrm{n})^H$, where $\mathbf{U}_\mathrm{n}$ is extracted from the singular value decomposition (SVD) of $\mathbf{R}^\mathrm{e}_{\mathrm{J_k}}$ [2],
  $$\mathbf{R}^\mathrm{e}_{\mathrm{J_k}} = [\mathbf{U}_\mathrm{s}\ \mathbf{U}_\mathrm{n}] \begin{bmatrix} \mathbf{\Lambda}_\mathrm{s} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Lambda}_\mathrm{n} \end{bmatrix} \begin{bmatrix} (\mathbf{U}_\mathrm{s})^H \\ (\mathbf{U}_\mathrm{n})^H \end{bmatrix}. \tag{4}$$
- During the preamble phase, because the jamming signal is nullified by multiplying the received signal in (1) with $\mathbf{W}_k$, BS-UE equivalent channel (i.e., $\mathbf{W}_k \mathbf{H}_k$) can be estimated using pilot signals and a channel estimator such as least-square (LS) or minimum mean-square error (MMSE) channel estimator.
- During the data transmission phase, which lasted for N_d samples, BS-UE communication is performed. The spectral efficiency of each BS-UE link is represented by
  $$C_k[n] = \log_2(1 + \delta_k[n]). \tag{5}$$
  where $\delta_k[n]$ is the received signal-to-interference-plus-noise ratio (SINR) at the kth UE during the nth frame.

B. Impact of Time-varying Correlations on Jamming Nullification Effectiveness

The beam-forming matrix $\mathbf{W}_k$ described in the previous subsection is derived from the left nullspace of the jamming channel. Therefore, under normal conditions, within jamming channel nullspace coherence time [10], $\mathbf{W}_k$ is capable of nullifying the jamming signal. However, as demonstrated in [7], when the correlations between transmitted jamming signals vary over time, they create a "virtual change" in the jamming channel, making $\mathbf{W}_k$ unable to suppress the jamming signal, even when the jamming channel does not physically changes. For brevity, we summarize the most important findings in [7] regarding the "virtual change" as follows.

Let ρ_{ij} be the correlation between the transmitted jamming signals from the ith and the jth jammer. Let ρ^e_{ij} and ρ^d_{ij} be ρ_{ij} values in the estimation phase and data transmission phase, respectively. Similarly, let $\mathbf{\Sigma}^\mathrm{e}_\mathrm{J}$ and $\mathbf{\Sigma}^\mathrm{d}_\mathrm{J}$ denote $\mathbf{\Sigma}_\mathrm{J}$ values in these two phases. The impact of the time-varying correlations on jamming suppression is given in [7] and demonstrated by Theorem 1 presented below.

Theorem 1: Let
$$\mathbf{\Sigma}^\mathrm{e}_\mathrm{J} = \mathbf{V}^\mathrm{e} \mathbf{S}^\mathrm{e} (\mathbf{V}^\mathrm{e})^H \text{ and } \mathbf{\Sigma}^\mathrm{d}_\mathrm{J} = \mathbf{V}^\mathrm{d} \mathbf{S}^\mathrm{d} (\mathbf{V}^\mathrm{d})^H$$
be the SVD of $\mathbf{\Sigma}^\mathrm{e}_\mathrm{J}$ and $\mathbf{\Sigma}^\mathrm{d}_\mathrm{J}$, respectively. Let $\mathbf{F}$ denotes the "virtual change" factor given by
$$\mathbf{F} = \mathbf{V}^\mathrm{d} \sqrt{\mathbf{S}^\mathrm{d} (\mathbf{S}^\mathrm{e})^{-1}} (\mathbf{V}^\mathrm{e})^H. \tag{6}$$
Then, the change over time from ρ^e_{ij} to ρ^d_{ij} causes a "virtual change" in the jamming channels from $\mathbf{Z}_k$ to $(\mathbf{Z}_k \mathbf{F})$.

Proof: The proof is given in [7]. ■

The interesting intuition of the impact of time-varying correlation on the jamming suppression can be found by examining the behavior of the "virtual change" factor $\mathbf{F}$. First, when the correlations are unchanged over time, we have $\mathbf{\Sigma}^\mathrm{d}_\mathrm{J} = \mathbf{\Sigma}^\mathrm{e}_\mathrm{J}$, $\mathbf{S}^\mathrm{d} = \mathbf{S}^\mathrm{e}, \mathbf{V}^\mathrm{d} = \mathbf{V}^\mathrm{e}$, and hence $\mathbf{F} = \mathbf{I}$. Therefore, there is no "virtual change" when the correlations are unchanged over time. For that, within the jamming channel nullspace coherence time, $\mathbf{W}_k$ derived from $\hat{\mathbf{G}}_k$ can be utilized to suppress the jamming signals when the correlations are unchanged over time, regardless of the correlation values. Second, when the correlations are time-varying, there is a non-identity "virtual change" factor $\mathbf{F}$, and the behavior of its element values are described by Corollary 1.1 below.

Corollary 1.1: When $|\rho^\mathrm{e}_{ij}| \to 1$ and $\rho^\mathrm{d}_{ij} \neq \rho^\mathrm{e}_{ij}$, the "virtual change" factor $\mathbf{F}$'s elements approach infinity.

Proof: The proof is given in [7]. ■

Therefore, from the UE receiver's observation, when the correlations vary over time, there is a "virtual change" $\mathbf{F}$ in the jamming channels. Distinctively, the elements of $\mathbf{F}$ approach infinity when $|\rho^\mathrm{e}_{ij}| \to 1$. As a result, $\hat{\mathbf{G}}_k$ becomes ineffective when the correlations are large and vary. In this case, using

$\hat{\mathbf{G}}_k$ to create $\mathbf{W}_k$ does not guarantee jamming nullification in the data transmission phase.

C. Jamming Signal Model

Given the above two observations on the behavior of the "virtual change" factor $\mathbf{F}$, we consider proactive jammers with the transmitted jamming signals designed to be resistant to the jamming nullification protocol given in Subsection III-A. Specifically, the correlations are time-varying and controlled by the jammers using the formula

$$\rho_{ij}(t) = \mathcal{J}(i,j,t), \forall i \# j \in \{1,2,...,N_J\}, \tag{7}$$

where $\mathcal{J}$ is a function unknown to the BS and the UEs. Note that the jammers can deliberately adjust the function $\mathcal{J}$ to make the jamming suppression even more challenging.

D. Problem Formulation

Given the BS-UEs communication system under jamming strategy described by equation (7), our objective is to continually optimize the length of the estimation phase and the data transmission phase (i.e., N_e and N_d, respectively) to achieve optimal system's spectral efficiency. Specifically, we tune the N_e and N_d values at each frame to achieve the optimal system's spectral efficiency because of the following reasons.

- First, N_e and N_d are optimzed to guarantee $\mathbf{W}_k$ being estimated when none of ρ^{e}_{ij} is closed to 1. As presented in Corollary 1.1, when $|\rho^{\mathrm{e}}_{ij}| \to 1$, the elements of the "virtual change" factor $\mathbf{F}$ approach infinity, resulting in a significant "virtual change", making the beam-forming matrix $\mathbf{W}_k$ to be unable to suppress the jamming signals.
- Second, by optimizing N_e and N_d, the system can avoid spending time monitoring the residual jamming signals as in [7] to update the beam-forming matrix. Hence, the system's spectral efficiency can be improved.
- Third, by varying N_d, the communication system can adapt to the change in the BS-UE channel condition. For example, when the channel coherence time decrease, the T_d values should be decreased to maintain an acceptable received SINR level. On the other hand, when the coherence time increases, the system can increase T_d to improve the communication phase percentage over the whole frame.

We mathematically formulate the problem as

$$\begin{aligned} \max_{N_e[n], N_d[n]} \quad & \sum_{n=1}^{N} \sum_{k=1}^{K} N_d[n] \log_2(1+\delta_k[n]) \\ \textbf{s.t.} \quad & N_e[n] \in \mathcal{N}_{\mathbf{e}} \\ & N_d[n] \in \mathcal{N}_{\mathbf{d}} \\ & \delta_k[n] \geq \delta_{min} \\ & \rho_{ij} \text{ controlled by } (7), \end{aligned} \tag{8}$$

where N is number of frames over a fix period of time, $\mathcal{N}_{\mathbf{e}} \triangleq \{N_e^1, N_e^2, ..., N_e^{L_e}\}$ and $\mathcal{N}_{\mathbf{d}} \triangleq \{N_d^1, N_d^2, ..., N_d^{L_d}\}$ are the set of L_e and L_d candidates for N_e and N_d, respectively, and δ_{min} is the required minimum SINR, below which the UE is considered to be outage.

The problem in (8) is a non-convex optimization problem because of the non-convexity of the first two constraints. More importantly, the jamming strategy, demonstrated by equation (2) is unknown to the BS and the UEs. To make the jamming suppression even more challenging, the jammers can adjust the function $\mathcal{J}$, making the old measurement data no longer representative of the current jamming strategy. To deal with such uncertainty and incomplete information, in the following, we describe the deep dueling Q-learning technique to solve the problem stated in (8).

IV. Deep Dueling Q-Learning for Jamming Suppression

A. Semi-Markov Decision Process (SMDP)

To maximize the long-term spectral efficiency, we use the semi-Markov decision process (SMDP) [11]. An SMDP is defined by a tuple $< t[n], \mathcal{S}, \mathcal{A}, r >$, where $t[n]$ is the nth decision epoch length, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, and r is the reward function. The SMDP allows the state transition to take place at irregular time steps. Therefore, the SMDP is more effective than the MDP in solving the optimization problem in (8), because we are optimizing the selections of N_e and N_d, which requires irregular state transition time.

1) State: There are several essential factors to consider for achieving the stated objective. The first factor is the received SINR levels of the UEs during the previous data communication phase. This is because the received SINR implicitly captures the BS-UE channel condition that affects the selection of N_d. The second factor, as demonstrated in the previous section, is the correlations between transmitted jamming signals during the estimation phase. This is because the correlations ρ^{e}_{ij} affect the magnitude of the "virtual change" factor $\mathbf{F}$, which directly affect the jamming nullification capability of $\mathbf{W}_k$. Therefore, the state space of the system can be defined as follows

$$\begin{aligned} \mathcal{S} \triangleq \big\{ [\delta_k, \rho^{\mathrm{e}}_{ij}] : & \forall k \in \{1,2,...,K\}, \\ & \forall i \# j; i,j \in \{1,2,...,N_J\} \big\}. \end{aligned} \tag{9}$$

2) Observation: In fact, the correlation coefficients ρ_{ij} between the transmitted jamming signals are controlled by the jammers (i.e., by formula (7)), and are unknown to the BS and the UEs. Moreover, the ρ^{e}_{ij} values are not directly observable by the BS nor the UEs. The ρ^{e}_{ij} values can merely be indirectly observed by examining the SVD of K received jamming signal covariance matrices $\mathbf{R}^{\mathrm{e}}_{\mathrm{J_k}}$ at the K UEs. In general, small correlations between transmitted jamming signals result in relatively equal singular values of $\mathbf{R}^{\mathrm{e}}_{\mathrm{J_k}}$, while large correlations result in the massive difference between the singular values.

Therefore, we formulate the problem as a partially observable MDP (POMDP) [11], where the state in (9) is replaced by the approximate state $\hat{\mathcal{S}}$ derived from the observations from the UEs. Specifically, the observation of the system is defined as:

$$\begin{aligned} \mathcal{O} \triangleq \big\{ (\delta_k, \bar{\Lambda}_l) : & \forall k \in \{1,2,...,K\}, \\ & \forall l \in \{1,2,...,N_J\} \big\}, \end{aligned} \tag{10}$$

where $\bar{\Lambda}_l = \frac{1}{K}\sum_{k=1}^{K}\Lambda_{kl}$, and Λ_{kl} is the lth largest singular value of $\mathbf{R}^{\mathrm{e}}_{\mathrm{J_k}}$. To generate the observation, K received SINR values and $(K * N_J)$ singular values Λ_{kl} are calculated. The observation is then obtained by concatenating K received SINR values and N_J average singular value $\hat{\Lambda}_l$.

We use the last H observations and actions as the approximate state, i.e., $\hat{s}[n] = \big[o[n], a[n-1], o[n-1], ..., a[n-H]\big]$. This formalism, referred to as the Hth-order history approach [11], generates a large but finite MDP in which each sequence is a distinct state [12]. As a result, we can apply standard reinforcement learning methods for MDPs to find the optimal action given the current approximate state [11], [12].

3) Action: The action space is defined as $\mathcal{A} \triangleq \{a : a \in \{1, 2, ..., L_e \times L_d\}\}$, where L_e and L_d are the dimension of $\mathcal{N}_{\mathbf{e}}$ and $\mathcal{N}_{\mathbf{d}}$, respectively, and

$$a = \begin{cases} 1, & N_e = N_e^1 \text{ and } N_d = N_d^1 \\ 2, & N_e = N_e^2 \text{ and } N_d = N_d^1 \\ ... & \\ L_e \times L_d, & N_e = N_e^{L_e} \text{ and } N_d = N_d^{L_d}. \end{cases}$$

4) Intermediate Reward: The intermediate reward is defined as the maximum achievable data transmitted during the data transmission phase, and zero if the received SINR during the data transmission phase is smaller than the minimum required received SINR. Specifically,

$$r[n] = \begin{cases} \sum_{k=1}^{K} N_d[n]\log_2(1+\delta_k[n]), & \delta_k[n] \geq \delta_{min} \\ 0, & \delta_k[n] < \delta_{min}. \end{cases} \quad (11)$$

5) Optimization Formulation: We target to obtain the optimal policy, denoted by π^*, that maximizes the average long-term reward [13] of the system, as represented in (8). Specifically, $\pi^* : \hat{\mathcal{S}} \rightarrow \mathcal{A}$ is a mapping from the observed approximate states to the actions taken by the BS. The optimization problem is then expressed as follows

$$\begin{aligned} \max_{\pi} \quad \mathcal{R}(\pi) &= \lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}(r[n]) \\ &= \lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\mathbb{E}(r(\hat{s}[n], \pi(\hat{s}[n]))), \end{aligned} \quad (12)$$

where $\mathcal{R}(\pi)$ is the long-term average reward of the system under the policy π.

B. Deep Dueling Q-learning Technique

The Q-learning [11] algorithm can help the BS transmitter find the optimal policy (i.e., a mapping from the state to the action) without requiring information about the jammers' strategy or channels conditions. However, as mentioned in the previous subsection, the state s is not fully observable by the BS and the UEs. Therefore, we use the approximate state $\hat{s}$ to derive the optimal policy. Nevertheless, the approximate state components (i.e., δ_k and $\bar{\Lambda}_l$) are continuous values, resulting in an infinite approximate state dimension. Quantization of the components can reduce the dimension. However, a smaller quantization step size (i.e., for better accuracy) results in a larger approximate state space, making the Q-learning algorithm to converges slowly. Moreover, the approximate state is composed of H latest observations and actions, further increases the approximate state dimension, and aggravates the slow-convergence issue. Therefore, we adopt the deep dueling Q-learning [14], which uses a neural network to efficiently obtain the optimal policy.

In particular, instead of finding and storing the optimal state-action value function $Q^*(\hat{s}, a)$ in a Q-table, a neural network is used as a nonlinear function approximator to estimate $Q^*(\hat{s}, a)$ value. Note that $Q(\hat{s}, a)$ is the expected discounted reward of the system starting from approximate state $\hat{s}$ selecting an action a, and $Q^*(\hat{s}, a)$ is the optimal value of $Q(\hat{s}, a)$. The input to the neural network is the approximate state $\hat{s}$, and the output of the neural network is the state-action values $Q^*(\hat{s}, a)$.

Let $\boldsymbol{\theta}$ be the parameters of the neural network, the problem of finding $Q^*(\hat{s}, a)$ becomes the problem of finding $\boldsymbol{\theta}^*$, which are the optimal values of $\boldsymbol{\theta}$. Accordingly, the state-action values function is now denoted by $Q(\hat{s}, a; \boldsymbol{\theta})$, and its optimal is denoted by $Q^*(\hat{s}, a; \boldsymbol{\theta}^*)$. The algorithm to iteratively optimize $\boldsymbol{\theta}$ is presented in Algorithm 1. It is based on the algorithm in [12], and formed by the following techniques.

- ϵ-greedy action selection policy: At each iteration in the training process, the agent implements *exploration* (by choosing a random action) with a probability of ϵ, or *exploitation* (by choosing the action that maximize the state-action value $Q(\hat{s}, a; \boldsymbol{\theta})$) with a probability of $1 - \epsilon$.
- Experience replay: Instead of using instant state-action value, the algorithm stores the transitions $(\hat{s}[i], a[i], r[i], \hat{s}[i+1])$ in a memory pool $\mathbf{D}$ of size D. The learning process is then performed based on random samples from $\mathbf{D}$. This technique allows the previous transitions to be used more than once, which improves training data efficiency. More importantly, by randomly selecting the training data from $\mathbf{D}$, the algorithm can remove the correlation between the consecutive training data.
- Target Q-network: This technique is performed by using a separate network, named target Q-network $\hat{\mathcal{Q}}$ for generating the target Q-values $y[j]$, as demonstrated in step 9 in Alg. 1. The target Q-network $\hat{\mathcal{Q}}$ is updated every C steps. In this way, the primary Q-network is slowly updated, which helps to reduce the correlations between the target and estimated Q-values, thereby improving the stability of the deep dueling Q-learning algorithm [12].
- Mini-batch gradient descent [15]: At each training iteration of the Q-learning algorithm, instead of performing gradient descent using the whole data memory $\mathbf{D}$, we randomly sample a mini-batch of size N_{mb} from $\mathbf{D}$, and then perform mini-batch gradient descent on the mini-batch training data. By setting $N_{mb} << D$, the training time can be reduced dramatically [15].

C. Deep Dueling Neural Network Structure

Unlike conventional recurrent neural networks (RNN) that have difficulty learning long-term dependencies of the inputs [16], the Long Short-Term Memory (LSTM) is capable of learning those dependencies, even with inputs consisting of

Algorithm 1 Deep Q-learning Based Jamming Suppression.

Initialize replay memory $\mathbf{D}$ with capacity D.
Initialize Q-network $\mathcal{Q}$ with random weights $\boldsymbol{\theta}$.
Initialize target Q-network $\hat{\mathcal{Q}}$ with weights $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}$.
for iteration $i = 1$ *to* I **do**
  Select action

$$a[i] = \begin{cases} \text{random action}, & \text{with probability } \epsilon \\ \arg\max_a Q(\hat{s}[i], a; \boldsymbol{\theta}), & \text{otherwise.} \end{cases}$$

  Perform $a[i]$, observe reward $r[i]$ and the next approximate state $\hat{s}[i+1]$.
  Store transition $(\hat{s}[i], a[i], r[i], \hat{s}[i+1])$ in $\mathbf{D}$.
  Sample random mini-batch of transitions $(s[j], a[j], r[j], s[j+1])$ from $\mathbf{D}$.
  Set $y[j] = r[j] + \gamma[i] \max_{a[j+1]} \hat{Q}(\hat{s}[j+1], a[j+1]; \hat{\boldsymbol{\theta}})$
  Perform mini-batch gradien descent [15] on $(y[j] - Q(\hat{s}[j], a[j]; \boldsymbol{\theta}))^2$ with respect to $\boldsymbol{\theta}$.
  Set $\hat{\mathcal{Q}} = \mathcal{Q}$ every C steps.
end for

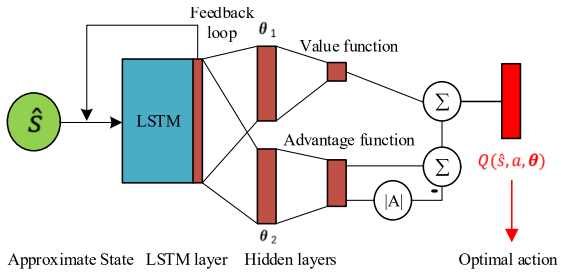

Fig. 2: LSTM-based deep dueling neural network.

more than 1000 discrete-time steps. That improvement is because the LSTM resolves the "vanishing gradients" and exploding gradients" [16], which are the main problems in the training process of the RNN. By using the LSTM, the network can capture the change in the correlations between jamming signals (i.e., by observing the sequential average singular values $\bar{\Lambda}_l$) and the change in channels condition (i.e., by observing the sequential received SINR δ_k). Fig. 2 illustrates the LSTM-based deep dueling neural network used in the proposed deep dueling Q-learning technique. First, the approximate state is used as the input to the LSTM layer. The LSTM layer learns the valuable information from the input (e.g., the correlations between transmitted jamming signals and the channels condition) and represents this information by the output of the LSTM. The output from the LSTM is then processed by two separated streams of fully connected hidden layers to calculate the values of states and the advantages of actions [17]. The values and the advantages are then used to generate $Q^*(\hat{s}, a; \boldsymbol{\theta}^*)$ at the output layer.

V. Simulation Results

To show the effectiveness of the proposed deep dueling Q-learning technique in nullifying jamming signals, we compare the following schemes, the history length is $H = 8$, unless otherwise specified.

- *Fix action*: The system uses a fixed pair of values for N_e and N_d. The performance metrics are calculated by averaging the performance of $(N_e \times N_d)$ action choices;
- *Technique in [7]*: The system uses the jamming nullification technique in [7], in which the residual jamming signals are measured, and the beam-forming matrix is updated whenever the residual exceeds a predefined value;
- *Proposed deep dueling Q-learning*: The values of N_e and N_d are determined by the optimal policy obtained using the proposed deep dueling Q-learning algorithm.

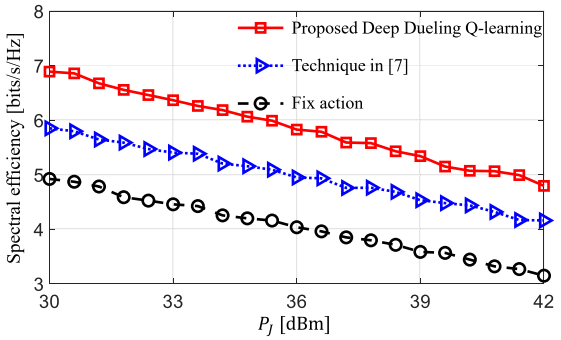

Fig. 3: Throughput for different jamming power level.

A. Spectral Efficiency Analysis

Fig. 3 shows the average spectral efficiency of each BS-UE communication link for different jamming nullification approaches for different values of the jamming power. For a fair comparison, the spectral efficiency of each technique is averaged over the UEs and normalized, taking into account the estimation phase time (because the system does not communicate during this phase) as

$$\frac{1}{NK} \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{N_d[n]}{N_d[n] + N_e[n]} \log_2(1 + \delta_k[n]). \tag{13}$$

As can be seen, the proposed deep dueling Q-learning achieved the highest spectral efficiency for all values of the jamming power, thanks to its ability to effectively adjust the N_e and N_d values according to the change in the correlations and channel conditions. On the other hand, the other two techniques have several limitations. While the technique in [7] spends an excessive amount of time monitoring the residual jamming signals and estimating the beam-forming matrix, thus reducing transmission time, the *fix action* technique cannot adapt to the change in the channel conditions, and more importantly, that in the correlations between transmitted jamming signals. Those limitations of [7] and *fix action* result in lower spectral efficiencies of the communication system.

B. Outage Probability Analysis

Fig. 4 illustrates the outage probability of the systems using three mentioned techniques for different values of the jamming power. As can be seen, the proposed deep dueling Q-learning technique and the techniques in [7] have very similar outage probabilities, and are much lower than that of the *fix action* technique. This is because both techniques effectively nullify the jamming signals. However, as aforementioned, the technique in [7] spends an excessive amount of time monitoring the residual jamming signals and estimating the beam-forming matrix, resulting in a lower spectral efficiency as aforementioned. On the other hand, the *fix action* technique cannot

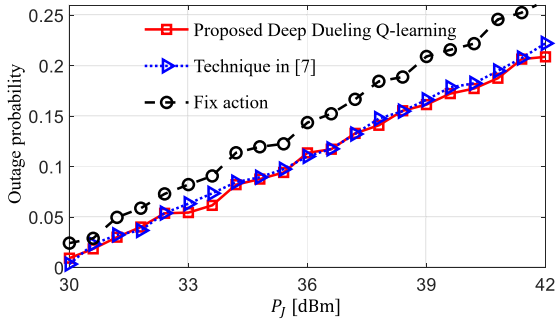

Fig. 4: Outage probability for different jamming power level.

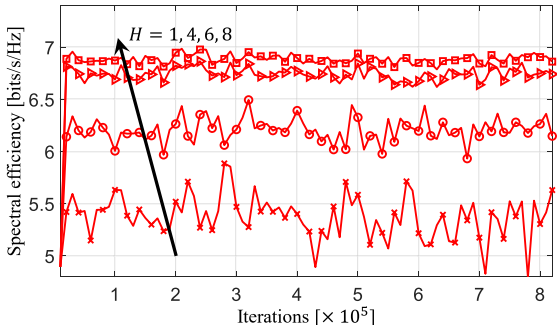

Fig. 5: Spectral efficiency convergence rate.

adapt to the change of the correlations between transmitted jamming signals and channels condition, resulting in many outage frames because of excessive residual jamming signals. Therefore, the proposed deep dueling Q-learning technique succeeds in increasing the system's spectral efficiency while keeping the outage probability at an acceptable level.

C. Impact of History Length H

Fig. 5 and Fig. 6 illustrate the impact of the historical length H on the convergence of the proposed deep dueling Q-learning technique. The jamming power used to generate these figures is $P_J = 30$ dBm. As can be seen, the deep dueling Q-learning algorithm converges after around 10^5 iterations. Moreover, a longer history length H results in a higher spectral efficiency and a lower outage probability. However, increasing the value of H also increases the computational complexity of the deep dueling Q-learning approach. As can be seen, the spectral efficiency and the outage probability do not dramatically improve as H increases from 6 to 8. Therefore, using $H = 6$ can balance the technique's performance and computational complexity.

VI. Conclusion

We have examined the impact of time-varying correlations between transmitted jamming signals on jamming nullification. We proposed the deep dueling Q-learning technique to effectively nullify jamming signals. Simulation results show that our techniques can achieve higher spectral efficiency and lower outage probability compared to the existing techniques.

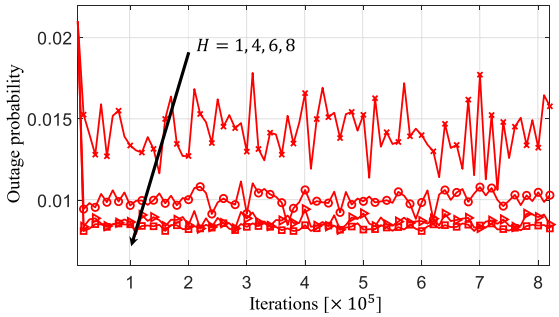

Fig. 6: Outage probability convergence rate.

References

[1] A. J. Fenn, *Adaptive antennas and phased arrays for radar and communications*, Boston, MA, USA, 2007.

[2] X. G. Doukopoulos and G. V. Moustakides, "Fast and stable subspace tracking," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1452–1465, 2008.

[3] T. T. Do, E. Björnson, E. G. Larsson, and S. M. Razavizadeh, "Jamming-resistant receivers for the massive MIMO uplink," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 1, pp. 210–223, 2017.

[4] Q. Yan, H. Zeng, T. Jiang, M. Li, W. Lou, and Y. T. Hou, "Jamming resilient communication using MIMO interference cancellation," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 7, pp. 1486–1499, 2016.

[5] M. H. Brady, M. Mohseni, and J. M. Cioffi, "Spatially-correlated jamming in gaussian multiple access and broadcast channels," in *Proc. CISS*, Princeton, NY, USA, Mar. 2006, pp. 1635–1639.

[6] J. Gao, S. A. Vorobyov, H. Jiang, and H. V. Poor, "Worst-case jamming on MIMO Gaussian channels," *IEEE Trans. Signal Process.*, vol. 63, no. 21, pp. 5821–5836, 2015.

[7] L. M. Hoang, J. A. Zhang, D. Nguyen, X. Huang, A. Kekirigoda, and K.-P. Hui, "Suppression of multiple spatially correlated jammers," *IEEE Trans. Veh. Tech-nol.*, 2021.

[8] M. K. Hanawal, D. N. Nguyen, and M. Krunz, "Jamming attack on in-band full-duplex communications: Detection and countermeasures," in *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, 2016, pp. 1–9.

[9] G. Strang, *Introduction to linear algebra*. Wellesley, MA, USA: Wellesley-Cambridge Press, 2016.

[10] A. Manolakos, Y. Noam, K. Dimou, and A. J. Goldsmith, "Blind null-space tracking for MIMO underlay cognitive radio networks," in *Proc. Global Commun. Conf.*, 2012, pp. 1223–1229.

[11] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. Cambridge, MA, USA: MIT Press, 2018.

[12] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[13] N. Van Huynh, D. T. Hoang, D. N. Nguyen, and E. Dutkiewicz, "Optimal and fast real-time resource slicing with deep dueling neural networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1455–1470, 2019.

[14] N. Van Huynh, D. N. Nguyen, D. T. Hoang, and E. Dutkiewicz, ""jam me if you can:" defeating jammer with deep dueling neural network architecture and ambient backscattering augmented communications," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 11, pp. 2603–2620, 2019.

[15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. Cambridge, MA, USA: MIT press, 2016.

[16] Y. Bengio, P. Frasconi, and P. Simard, "The problem of learning long-term dependencies in recurrent networks," in *Proc. IEEE Int. Conf. Neural Netw.* IEEE, 1993, pp. 1183–1188.

[17] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," Nov 2015. [Online]. Available: https://arxiv.org/abs/1511.06581