Two-dimensional Anti-jamming Mobile Communication Based on Reinforcement Learning

Liang Xiao, *Senior Member, IEEE*, Donghua Jiang, Dongjin Xu, Hongzi Zhu, *Member, IEEE*, Yanyong Zhang, *Fellow, IEEE*, and H. Vincent Poor, *Fellow, IEEE*,

***Abstract*—By using smart radio devices, a jammer can dynamically change its jamming policy based on opposing security mechanisms; it can even induce the mobile device to enter a specific communication mode and then launch the jamming policy accordingly. On the other hand, mobile devices can exploit spread spectrum and user mobility to address both jamming and interference. In this paper, a two-dimensional anti-jamming mobile communication scheme is proposed in which a mobile device leaves a heavily jammed/interfered-with frequency or area. It is shown that, by applying reinforcement learning techniques, a mobile device can achieve an optimal communication policy without the need to know the jamming and interference model and the radio channel model in a dynamic game framework. More specifically, a hotbooting deep Q-network based two-dimensional mobile communication scheme is proposed that exploits experiences in similar scenarios to reduce the exploration time at the beginning of the game, and applies deep convolutional neural network and macro-action techniques to accelerate learning in dynamic situations. Several real-world scenarios are simulated to evaluate the proposed method. These simulation results show that our proposed scheme can improve both the signal-to-interference-plus-noise ratio of the signals and the utility of the mobile devices against cooperative jamming compared with benchmark schemes.**

***Index Terms*—Mobile devices, jamming, reinforcement learning, game theory, deep Q-network.**

I. Introduction

By injecting faked or replayed signals, a jammer aims to interrupt the ongoing communication of mobile devices such as smartphones, laptops and mobile sensing robots, and even result in denial of service (DoS) attacks in wireless networks [1]–[5]. With the pervasion of smart radio devices such as universal software radio peripherals (USRPs), smart jammers can cooperatively and flexibly choose their jamming policies to block the mobile devices efficiently [6], [7]. Jammers can even induce the mobile device to enter a specific communication mode and then launch the jamming attacks accordingly.

Radio devices usually apply spread spectrum techniques, such as frequency hopping and direct-sequence spread spectrum to address jamming attacks [8]. However, if most frequency channels in the receiver location are blocked by jammers and/or strongly interfered by electric appliances such as microwaves and other communication radio devices, spread spectrum alone can't improve the communication performance such as the signal-to-interference-plus-noise ratio (SINR) of the received signals and the bit error rate (BER) of the messages.

Therefore, we develop a two-dimensional (2-D) anti-jamming mobile communication system that applies both frequency hopping and user mobility to address jamming and interference. In this system, a mobile device will move to another location for better communication efficiency if the current location is severely jammed or interfered. This system has to make a tradeoff between the communication efficiency and the cost due to the change of the geographical location before finishing the communication task as well as the switch of the frequency channel. Mobile devices as secondary users in cognitive radio networks (CRNs) have to avoid interfering with the ongoing communication of primary users (PUs).

In this work, we formulate the repeated interactions between a mobile device using the two-dimensional anti-jamming communication scheme and jammers as a non-zero-sum dynamic anti-jamming communication game as the mobile device aims to improve its communication performance such as the SINR of the signals with less transmission cost while the jammers are concerned with the jamming cost. The communication decisions of the mobile device in the dynamic game can be formulated as a Markov decision process (MDP). Therefore, reinforcement learning (RL) techniques such as Q-learning can be used by mobile devices to achieve an optimal communication policy via trail-and-error without being aware of the jamming and network model [9]. We have developed a Q-learning based 2-D anti-jamming mobile communication scheme in [10] to choose the transmit power and determine whether to leave the location against jamming and strong interference. However, the Q-learning based 2-D mobile communication scheme suffers from the curse of high-dimensionality, i.e., the learning speed is extra slow, if the mobile device has a large number of frequency channels and can observe a large range of the feasible SINR levels. In this work, deep Q-network (DQN) as a deep reinforcement learning technique is used to accelerate the

This work was supported in part by the National Natural Science Foundation of China under Grant 61671396 and 91638204, in part by the U.S. National Science Foundation under Grants CNS-1702808 and ECCS-1647198, and in part by the open research fund of the National Mobile Communications Research Laboratory, Southeast University (No.2018D08).

L. Xiao, D. Jiang and D. Xu are with the Dept. of Communication Engineering, Xiamen University, Xiamen 361005, China (e-mail: lxiao@xmu.edu.cn). L. Xiao is also with the National Mobile Communications Research Laboratory, Southeast University, China.

H. Zhu is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, China. E-mail: hongzi@sjtu.edu.cn.

Y. Zhang is with the Wireless Information Networks Laboratory (WINLAB), Rutgers University, North Brunswick, USA. E-mail: yyzhang@winlab.rutgers.edu.

H. Vincent Poor is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544. E-mail: poor@princeton.edu.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2856854, IEEE Transactions on Vehicular Technology

learning of the mobile communication system for the case with a large number of frequency channels and jamming strengths. More specifically, a mobile device uses a deep convolutional neural network (CNN) to compress the state space consisting of the previous communication performance and jamming strength and thus improves the communication performance against jamming and strong interference.

We design a fast DQN based communication system that applies the macro-action technique as presented in [11] to further improve the learning speed. This scheme combines the power allocation and mobility decisions in a number of time slots as macro-actions and explores their quality values as a whole. The hotbooting technique as a transfer learning method is applied to exploit the previous anti-jamming communication experiences in similar scenarios to initialize the learning parameters such as the CNN weights. This technique helps mobile devices save the random exploration at the initial learning stage to resist jamming attacks.

This scheme can be implemented in three mobile applications against jammers and interference sources: (1) The command dissemination of a mobile server to devices such as smart TVs against jamming and interference, (2) The sensing report transmission of a mobile sensing robot to a server via several access points (APs), and (3) The sensing report transmission against two mobile jammers that randomly change their locations. Simulation results show that our proposed mobile communication scheme outperforms the benchmark mobile communication based scheme as developed in [10] with a faster learning speed, a higher SINR of the signals and a higher utility.

The main contributions of this paper are summarized as follows:

- We provide a frequency-spatial 2-D anti-jamming mobile communication scheme to resist jamming and interference and formulate a non-zero-sum dynamic game for the anti-jamming mobile communications.
- We implement the communication scheme in the command dissemination of a mobile server to radio devices and the sensing report transmission of a mobile sensing robot against both jamming and interference.
- We propose a fast DQN based 2-D mobile communication algorithm that applies DQN, macro-actions and hotbooting techniques to achieve the optimal frequency selection and mobility strategy without being aware of the jamming and network model. This algorithm accelerates learning and improves the communication performance compared with the benchmark Q-learning based and the DQN based communications in [10].

The rest of this paper is organized as follows. We review related work in Section II and present the system model in Section III. We propose a fast DQN based communication system in Section IV. We provide simulation results in Section VI and conclude this work in Section VII.

II. Related Work

Game theory has been applied to study the power allocation of the anti-jamming in wireless communication. For instance, the Colonel Blotto anti-jamming game presented in [12] provides a power allocation strategy to improve the worst-case performance against jamming in cognitive radio networks. The power control Stackelberg game as presented in [13] formulates the interactions among a source node, a relay node and a jammer that choose their transmit power in sequence without interfering with primary users. The transmission Stackelberg game developed in [14] helps build a power allocation strategy to maximize the SINR of signals in wireless networks. The prospect-theory based dynamic game in [15] investigates the impact of the subjective decision making process of a smart jammer in cognitive networks under uncertainties. The stochastic game formulated in [16] investigates the power allocation of a user against a jammer under uncertain channel power gains.

Game theory has been used for providing insights on the frequency channel selection against jamming. For instance, the stochastic channel access game investigated in [17] helps a user to choose the control channel and the data channel to maximize the throughput against jamming. The Bayesian communication game in [18] studies the channel selection against smart jammers with unknown types of intelligence. The zero-sum game as proposed in [19] investigates the frequency hopping and the transmission rate control to improve the average throughput against jamming. The game-theoretical anti-jamming channel selection scheme as developed in [20] increases the payoffs of mobile users and improves the communication performance against jamming.

Reinforcement learning techniques enable an agent to achieve an optimal policy via trials in Markov decision process. The Q-learning based power control strategy developed in [13] makes a tradeoff between the defense cost and the communication efficiency without being aware of the jamming model. The Q-learning based channel allocation scheme as proposed in [21] can achieve an optimal channel access strategy for a radio transmitter with multiple channels in the dynamic game. The synchronous channel allocation in [22] applies Q-learning to proactively avoid using the blocked channels in cognitive radio networks. The WoLF-Q based anti-jamming communication strategy as proposed in [23] selects transmit channel ID and the transmit power to resist sweeping jamming. An anti-jamming communication scheme as developed in [24] uses the state-action-reward-action-state-action method to choose the transmit channel to increase the payoff against jamming compared with Minimax-Q. The multi-agent reinforcement learning (MARL) based channel allocation as proposed in [25], [26] enhances the transmission and sensing capabilities for cognitive radio users. The MARL based power control strategy as developed in [27] accelerates the learning of the energy harvesting communication system against intelligent adversaries.

The 2-D anti-jamming mobile communication system proposed in [10] uses both frequency and spatial diverting to improve the communication performance against jamming and applies DQN to derive an optimal policy without knowing the jamming and interference model and the radio channel model.

In this work, we present a fast DQN based power and mobile control scheme that applies the hotbooting and macro-actions techniques to accelerate learning and thus improve the jamming resistance of the communication scheme as proposed in [10] for the mobile communication system with a large number of channels. We investigate the applications of this scheme in the sensing report transmission of a mobile sensing robot and the command dissemination of a mobile server to the smart devices against jamming and interference. We evaluate the performance of our proposed schemes against both static and mobile jammers in the sensing report transmission.

III. System Model

A. Network Model

A mobile device such as a smartphone and a mobile sensing robot aims to transmit messages over N frequency channels to serving radio nodes such as an AP or smart devices against jamming. All the radio nodes are assumed to share a frequency pattern set denoted by $C = [C_\psi]_{1\leq\psi\leq\vartheta}$ before the transmission, where ϑ is the size of the frequency pattern set and the ψ-th frequency pattern C_ψ consists of the channel indexes used by the mobile device and the receiver during κ time slots with $C_\psi = \left[c_\psi^{(i)}\right]_{1\leq i\leq\kappa}$. The mobile device sends a message to the target receiver at time k at channel $f^{(k)} = c_\psi^{k \bmod \kappa\ +1}$.

TABLE I: SUMMARY OF SYMBOLS AND NOTATIONS

Notation	Description
N	Number of frequency channels
J	Number of jammers
P_s	Transmit power of the mobile device
$P_{J/f}$	Jamming/interference power
$f^{(k)}$	The chosen channel at time k
$\psi^{(k)}$	The chosen frequency pattern index
$\phi^{(k)}$	User mobility indicator
P	The maximum transmit power
ϑ	Number of frequency patterns
κ	Length of a frequency pattern
N_J	Number of channels for sweep jammer
N_r	Number of channels for reactive jammer
$\mathbf{h}_s^{(k)}$	Channel power gains of the mobile device
$\mathbf{h}_j^{(k)}$	Channel power gains of jammer
$\lambda^{(k)}$	Absence of PU at time k
$\eta^{(k)}$	Status of the interference source
σ	Receiver noise power
C_m	Cost of user mobility
C_h	Cost of frequency hopping
C_p	Unit transmission cost
$u^{(k)}$	Utility of the mobile device
$\mathbf{s}^{(k)}$	System state
ξ	Number of the SINR quantization levels
γ	Discount factor in the learning algorithm
W	Size of the state-action pairs in the CNN
$\boldsymbol{\varphi}^{(k)}$	State sequence at time k
$\boldsymbol{\theta}^{(k)}$	CNN weights at time k
$\mathbf{e}^{(k)}$	Experience at time k
α	Learning rate
B	Size of the CNN minibatch
$\mathcal{M}$	Macro-actions set
Φ	Number of the macro-actions
ζ	Length of a macro-action
p	Jammer mobility probability

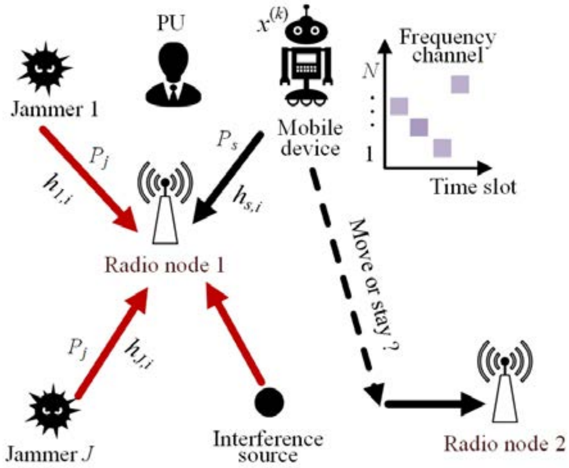


Fig. 1: Network model of the 2-D anti-jamming communication of a mobile device with N frequency channels, against J jammers and interference sources.

As shown in Fig. 1, the mobile device chooses the transmit power denoted by $P_s^{(k)}$ and whether to move its location denoted by $\phi^{(k)}$ at time k. The feasible transmit power $P_s^{(k)} \leq P$ is quantized into $L+1$ levels, where P is the maximum transmit power. The mobile device stays in the same location if $\phi^{(k)} = 0$; and it moves geographically to connect to a new radio node if otherwise. The mobile device has to avoid interfering with the local PUs and address the interference sources nearby.

Upon receiving the message, the serving radio node evaluates the BER of the message to estimate the SINR of the signals and quantizes the SINR into ξ levels. The radio node also chooses the frequency pattern index $\psi^{(k)}$ and sends the SINR and $\psi^{(k)}$ to the mobile device on the feedback channel.

The mobile device has to avoid interfering with the communication of the PU if in a cognitive radio network. The absence of the PU is denoted by $\lambda^{(k)}$, which equals 0 if the mobile device detects a PU accessing channel $f^{(k)}$ in the location and 1 otherwise. The mobile device applies a spectrum sensing technique, such as the energy detection as presented in [28] to detect the PU presence and thus obtains $\lambda^{(k)}$. Let the channel vector $\mathbf{h}_s^{(k)} = \left[h_{s,i}^{(k)}\right]_{1\leq i\leq N}$ denote the channel power gains of the N channels from the mobile device to the serving radio node, C_h be the cost of frequency hopping to the mobile device, C_p be the unit transmission cost and C_m be the extra cost of user mobility.

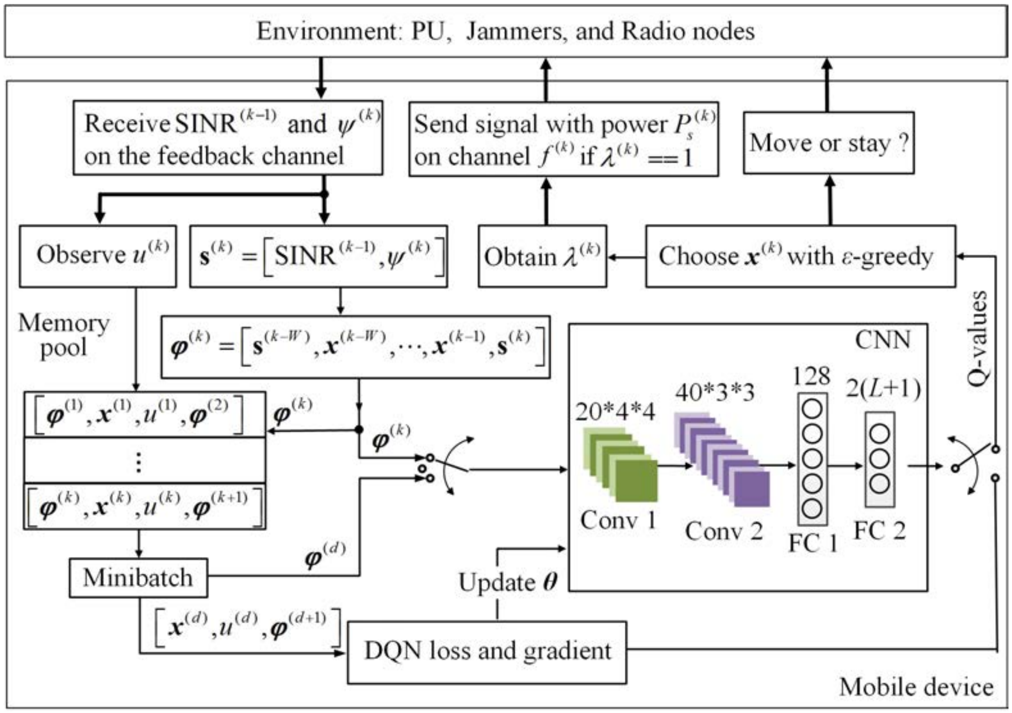

Environment: PU, Jammers, and Radio nodes
Receive SINR$^{(k-1)}$ and $\psi^{(k)}$ on the feedback channel
Send signal with power $P_s^{(k)}$ on channel $f^{(k)}$ if $\lambda^{(k)} == 1$
Move or stay ?
Observe $u^{(k)}$
$\mathbf{s}^{(k)} = [\text{SINR}^{(k-1)}, \psi^{(k)}]$
Obtain $\lambda^{(k)}$
Choose $\boldsymbol{x}^{(k)}$ with ε-greedy
Memory pool
$\boldsymbol{\varphi}^{(k)} = [\mathbf{s}^{(k-W)}, \boldsymbol{x}^{(k-W)}, \cdots, \boldsymbol{x}^{(k-1)}, \mathbf{s}^{(k)}]$
$[\boldsymbol{\varphi}^{(1)}, \boldsymbol{x}^{(1)}, u^{(1)}, \boldsymbol{\varphi}^{(2)}]$
⋮
$[\boldsymbol{\varphi}^{(k)}, \boldsymbol{x}^{(k)}, u^{(k)}, \boldsymbol{\varphi}^{(k+1)}]$
$\boldsymbol{\varphi}^{(k)}$
$\boldsymbol{\varphi}^{(k)}$
$\boldsymbol{\varphi}^{(d)}$
Minibatch
$[\boldsymbol{x}^{(d)}, u^{(d)}, \boldsymbol{\varphi}^{(d+1)}]$
CNN
20*4*4
40*3*3
128
2(L+1)
Conv 1
Conv 2
FC 1
FC 2
Q-values
Update $\boldsymbol{\theta}$
DQN loss and gradient
Mobile device

Fig. 2: DQN based 2-D anti-jamming mobile communication scheme.

B. Jamming Model

A jammer sends replayed or faked signals with power $P_j^{(k)} \leq P_J$ on the selected jamming channels to interrupt the ongoing communication of the mobile device, where P_J is the maximum jamming power. If failing to do that, the jammer also aims to reduce the SINR of the signals received by the radio node with less jamming power. We will consider four types of jamming attacks similar to [29]:

- A random jammer with power P_J randomly selects a jamming channel in each time slot, using the same jamming channel with a probability $1-\epsilon$ and a new channel with a probability ϵ.
- A sweep jammer blocks N_J neighboring channels in each time slot from the N channels in sequence and each channel is jammed with jamming power P_J/N_J.
- A reactive jammer as the most harmful chooses the jamming policy based on the the ongoing communication. The jammer detects radio power over N_r channels and sends jamming signals on the active channels with the jamming power $P_j^{(k)}$ that is chosen to maximize the jamming utility $u_j^{(k)}$ given by

$$u_j^{(k)} = \widehat{\text{SINR}}^{(k)} - C_j P_j^{(k)}, \tag{1}$$

  where C_j is the jamming cost.
- A mobile jammer changes its geographic location.

The jamming channel chosen by jammer j at time k is denoted by $y_j^{(k)} \in [1, \cdots, N]$. For simplicity, we define the action set of the J jammers at different locations in the area $\mathbf{y}^{(k)} = \left[y_j^{(k)}\right]_{1\leq j\leq J}$. By applying smart and programable radio devices, these jammers sometimes can block all the radio channels if the serving node is close enough to the jammer.

The status of the interference source at time k is denoted by $\eta^{(k)}$, which equals 1 if it interferes the ongoing message transmission of the mobile device with power P_f and 0 otherwise. The receiving noise power is denoted by σ. The channel power gains from the J jammers to the serving radio node on the N channels are denoted by $\mathbf{h}_j^{(k)} = \left[h_{j,i}^{(k)}\right]_{1\leq j\leq J, 1\leq i\leq N}$. Some interference sources and mobile jammers can block the data transmission from the mobile device in the new location to the new radio node. On the other hand, the new link is not impacted by the static jammers and weak interference sources in the previous location due to the large path-loss fading. For ease of reference, important notations are summarized in Table 1

IV. Fast DQN Based 2-D anti-jamming mobile communication scheme

The repeated interactions between the mobile device and the jammer are formulated as a non-zero-sum dynamic game, in which the communication scheme of the mobile device can be viewed as a MDP, as the future state observed by the mobile device is independent of the previous system state and action given the current state and communication scheme. Without being aware of the jamming and interference model and the radio channel model, a mobile device can apply reinforcement learning techniques such as Q-learning to achieve an optimal communication policy via trial-and-error in the dynamic game.

The learning speed of the Q-learning based 2-D communication algorithm proposed in our previous work in [10] suffers from the curse of high-dimensionality, i.e., the required convergence time increases with the dimension of the state space and

Algorithm 1: Fast DQN based 2-D mobile communication

```
1  $\boldsymbol{\theta}^{(0)} \leftarrow \boldsymbol{\theta}_*$
2  Initialize $\gamma$, $T$, $\mathcal{D}=\emptyset$, $\zeta$, $\Phi$, $\mathcal{M}=\emptyset$, $\mathrm{SINR}^{(0)}$, $\psi^{(1)}$ and $\bar{\mathbf{u}}=\mathbf{0}$
3  for $t=1,\cdots,T$ do
4      Choose $\boldsymbol{x}^{(t)}=\left[P_s^{(t)},\phi^{(t)}\right]$ at random
5      if $\phi^{(k)}$ == 1 then
6          Change the location and connect to a new radio node
7      Observe the absence of PU and set $\lambda^{(k)}$
8      if $\lambda^{(k)}$ == 0 then
9          Keep silence
10     else
11         $f^{(k)} \leftarrow c_{\psi}^{k \bmod \kappa+1}$
12         Send signals on channel $f^{(k)}$ with power $P_s^{(k)}$
13     end
14     Receive the $\mathrm{SINR}^{(k)}$ and $\psi^{(k+1)}$ on the feedback channel
15     Obtain $u^{(k)}$ and $\mathbf{s}^{(k+1)}=\left[\mathrm{SINR}^{(k)},\psi^{(k+1)}\right]$
16     $\bar{u}\left(\boldsymbol{x}^{(t)}\right) \leftarrow \max\left\{u^{(t)},\bar{u}\left(\boldsymbol{x}^{(t)}\right)\right\}$
17 end
18 $A=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{2L+2}\}$, $\forall\bar{u}(\boldsymbol{x}_i)>\bar{u}(\boldsymbol{x}_j)$, $i<j$
19 Store $m_i=\left\{\boldsymbol{x}_i^{(1)},\cdots,\boldsymbol{x}_i^{(\zeta)}\right\},\forall 1\le i\le\Phi$ in $\mathcal{M}$
20 $\mathbf{s}^{(1)}=\left[\mathrm{SINR}^{(0)},\psi^{(1)}\right]$
21 for $k=1,2,\cdots$ do
22     if $k\le W$ then
23         Choose $\boldsymbol{x}^{(k)}=\left[P_s^{(k)},\phi^{(k)}\right]$ at random
24     else
25         Obtain the CNN outputs with the input $\boldsymbol{\varphi}^{(k)}$
26         Choose $\boldsymbol{x}^{(k)}$ via (4)
27     end
28     Perform Steps 5-7
29     if $\lambda^{(k)}$ == 0 then
30         Keep silence
31     else
32         if $\boldsymbol{x}^{(k)}\in\mathcal{M}$ then
33             Follow the macro-action $\boldsymbol{x}^{(k)}$ in the next $\zeta$ time
               slots
34             Obtain $\left[u^{(v)}\right]_{k\le v\le k+\zeta-1}$ and observe a series of
               states $\left[\mathbf{s}^{(l)}\right]_{k+1\le l\le k+\zeta}$
35             Obtain a series of the CNN inputs
               $\left[\boldsymbol{\varphi}^{(i)}\right]_{k+1\le i\le k+\zeta}$
36             Calculate $U^{(k)}$ via (6)
37             $\mathcal{D}\leftarrow\left\{\boldsymbol{\varphi}^{(k)},\boldsymbol{x}^{(k)},U^{(k)},\boldsymbol{\varphi}^{(k+\zeta)}\right\}\cup\mathcal{D}$
38             $k\leftarrow k+\zeta$
39         else
40             Perform Steps 11-15 and 37
41         end
42     end
43     for $d=1,2,\cdots,B$ do
44         Select $\left(\boldsymbol{\varphi}^{(d)},\boldsymbol{x}^{(d)},u^{(d)},\boldsymbol{\varphi}^{(d+1)}\right)\in\mathcal{D}$ at random
45         if $\boldsymbol{x}^{(d)}\in\mathcal{M}$ then
46             Calculate $R$ via (5)
47         else
48             Calculate $R$ via (8)
49         end
50     end
51     Update $\boldsymbol{\theta}^{(k)}$ via (9)
52 end
```

the feasible communication strategy set, which increases with the number of frequency channels and the power quantization levels used by the mobile device. Therefore, we proposed a 2-D mobile communication scheme based on the deep Q-network, a deep reinforcement learning technique that applies deep convolutional neural networks to compress the state space observed by the mobile device.

Upon receiving the feedback from the radio node, the mobile device extracts the estimated SINR and the frequency pattern index. The mobile device detects the presence of PUs $\psi^{(k)}$, and formulates the state as $\mathbf{s}^{(k)}=\left[\mathrm{SINR}^{(k-1)},\psi^{(k)}\right]\in\mathbf{S}$, where $\mathbf{S}$ is the state set, whose dimension is $|\mathbf{S}|=\vartheta\xi$. The mobile device applying the reinforcement learning chooses transmit power $P_s^{(k)}$ and determines whether to change the location $\phi^{(k)}$, with the communication strategy denoted by $\boldsymbol{x}^{(k)}=\left[P_s^{(k)},\phi^{(k)}\right]\in\mathcal{X}$, where $\mathcal{X}$ is the action space.

Upon sending a message, the mobile device evaluates the SINR from the feedback information sent by the radio node and computes the utility received in this time slot based on both the communication performance criteria such as the SINR of the signals and the communication cost including the channel hopping overhead and the mobility overhead, i. e.,

$$u^{(k)}=\widehat{\mathrm{SINR}}^{(k)}-C_pP_s^{(k)}-C_m\phi^{(k)} \\ -C_h\mathcal{F}\left(f^{(k)}-f^{(k-1)}\right), \qquad (2)$$

where $\mathcal{F}(\varsigma)$ is an indicator function that equals 0 if ς equals 0, and 1 otherwise. The utility evaluation enables the mobile device to make a tradeoff between the communication performance and the cost against jamming.

As illustrated in Fig. 2, the communication strategy of the mobile device is chosen based on the quality function or Q-function of the current system state, which is the expected discounted long-term reward for each state-strategy pair, and defined as

$$Q(\mathbf{s},\boldsymbol{x})=\mathbb{E}_{\mathbf{s}'\in\mathbf{S}}\left[u^{(k)}+\gamma\max_{\boldsymbol{x}'\in\mathcal{X}}Q\left(\mathbf{s}',\boldsymbol{x}'\right)\middle|\mathbf{s},\boldsymbol{x}\right], \qquad (3)$$

where $\mathbf{s}'$ is the next state if the mobile device takes strategy $\boldsymbol{x}$ at state $\mathbf{s}$, and the discount factor γ represents the uncertainty of the mobile device regarding the future reward in the dynamic game against jamming and interference.

The deep convolutional neural network is a nonlinear function approximator to evaluate the Q-value in (3) for each communication policy against jamming, since the state set size $|\mathbf{S}|$ is too large for a Q-learning based scheme to quickly achieve an optimal policy. This deep RL based communication scheme compresses the state space that the mobile device observes into a small feature space. The CNN outputs are the basis to choose the communication channel and the mobility suggestion.

The state sequence at time k denoted by $\boldsymbol{\varphi}^{(k)}$ consists of the current system state and the previous W system state-strategy pairs, i.e., $\boldsymbol{\varphi}^{(k)}=\left[\mathbf{s}^{(k-W)},\boldsymbol{x}^{(k-W)},\cdots,\boldsymbol{x}^{(k-1)},\mathbf{s}^{(k)}\right]$. The size of the system state-strategy pairs W is set to make a tradeoff between the memory requirements and the anti-jamming communication performance. The memory overhead of the mobile device slightly increases with the size of the system state-strategy pairs, since the memory pool only stores the latest related experiences to save memory space. As shown in Fig. 2, the state sequence $\boldsymbol{\varphi}^{(k)}$ is reshaped into a $N_C \times N_C$ matrix and taken as the input to the CNN.

As shown in Fig. 2, the CNN consists of two convolutional (Conv) layers and two fully connected (FC) layers. The first Conv layer includes F_1 filters, each with size $N_1 \times N_1$ and stride n_1. The second Conv layer has F_2 filters, each with size $N_2 \times N_2$ and stride n_2. Both layers use the rectified linear units (ReLU) as the activation function. The first FC layer involves F_3 rectified linear units, and the second FC layer has $2(L+1)$ outputs for each feasible strategy. The filter weights of the four layers in the CNN at time k are denoted by $\boldsymbol{\theta}^{(k)}$, which are updated at each time slot based on the experience replay. The output of the CNN is used for estimating the values of the Q-function for the $2(L+1)$ actions, $Q\left(\boldsymbol{\varphi}^{(k)}, \boldsymbol{x} | \boldsymbol{\theta}^{(k)}\right), \forall \boldsymbol{x} \in \mathcal{X}$.

The communication policy $\boldsymbol{x}^{(k)}$ is chosen based on the ϵ-greedy algorithm to avoid staying in the local maximum. For example, such an algorithm helps the mobile device change its location and connect to a new serving radio node if the feedback channel is jammed. More specifically, the optimal communication policy with the highest Q-value is chosen with a high probability $1-\epsilon$, and other feasible strategies are chosen with a small probability, i.e.,

$$\Pr\left(\boldsymbol{x}^{(k)} = \dot{\boldsymbol{x}}\right) = \begin{cases} 1-\epsilon, & \dot{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}' \in \mathcal{X}} Q\left(\boldsymbol{\varphi}^{(k)}, \boldsymbol{x}'\right) \\ \frac{\epsilon}{2L+1}, & \text{o.w.} \end{cases} \tag{4}$$

The hotbooting process as presented in [3] exploits the previous anti-jamming communication experiences in I similar communication scenarios each lasting K time slots to initialize the filter weights of the CNN as $\boldsymbol{\theta}^{*}$. The temporal abstraction accelerates the learning for the large action space, which takes hierarchical multi-step actions as macro-actions or macros at different timescales. The macros are deterministic sequences of the power allocation and mobility decisions, i.e., a macro-action $m = \left[\boldsymbol{x}^1, \cdots, \boldsymbol{x}^{\zeta}\right] \in \mathcal{M}$, where $\mathcal{M}$ is the set of all macros and ζ is the length of a macro-action.

The mobile device transmits a message with a randomly chosen communication strategy $\boldsymbol{x}$ and evaluates the SINR and the utility. All the communication strategy experiences are sorted according to the utility. The top Φ communication strategies are chosen to construct the macros. Each macro-action m consists of the same strategy in ζ time slots in sequence.

Once a macro-action is chosen, the mobile device will transmit the message by following the communication strategy sequence which is predefined by the macro-action, observe a series of states $\left[\mathbf{s}^{(l)}\right]_{k+1 \le l \le k+\zeta}$ and evaluate the utility sequence $\left[u^{(v)}\right]_{k \le v \le k+\zeta-1}$. The optimal target Q-function in the fast DQN has to include the macros and is updated according to the cumulative discounted reward [11]. More specifically, during a multi-step transition from state $\mathbf{s}^{(k)}$ to state $\mathbf{s}^{(k+\zeta)}$ with macro-action m, the approximate optimal target Q-function with macros is updated by

$$R = U^{(k)} + \gamma^{\zeta} \max_{\boldsymbol{x}' \in \mathcal{X}} Q\left(\boldsymbol{\varphi}^{(k+\zeta)}, \boldsymbol{x}'; \boldsymbol{\theta}^{(k-1)}\right), \tag{5}$$

where $U^{(k)}$ is the cumulative discounted reward defined as

$$U^{(k)} = \sum_{i=0}^{\zeta-1} \gamma^i u^{k+i}. \tag{6}$$

After applying macros, the mobile device updates the number of the CNN outputs to $2(L+1)+\Phi$.

As summarized in Algorithm 1, the mobile device observes the SINR of the signals from the serving radio node at time k to update the system state and receives utility $u^{(k)}$. According to the next state sequence $\boldsymbol{\varphi}^{(k+1)}$, the new experience $\mathbf{e}^{(k)} = \left\{\boldsymbol{\varphi}^{(k)}, \boldsymbol{x}^{(k)}, u^{(k)}, \boldsymbol{\varphi}^{(k+1)}\right\}$ is stored in the memory pool $\mathcal{D} = \left\{\mathbf{e}^{(1)}, \cdots, \mathbf{e}^{(k)}\right\}$. By applying the experience replay, the mobile device chooses an experience $\mathbf{e}^{(d)}$ from the memory pool $\mathcal{D}$ at random, with $1 \le d \le k$ to update $\boldsymbol{\theta}^{(k)}$. By applying the stochastic gradient descent (SGD) algorithm, this scheme samples a subset of the loss functions at every step to reduce the computational complexity compared with the gradient descent algorithm. The stochastic nature of the SGD algorithm also avoids staying in the local minima in the learning process. This scheme minimizes the mean-squared error of the target optimal Q-function value and uses the minibatch updates for the loss function chosen by [10] as

$$L\left(\boldsymbol{\theta}^{(k)}\right) = \mathbb{E}_{\boldsymbol{\varphi}, \boldsymbol{x}, u, \boldsymbol{\varphi}'}\left[\left(R - Q\left(\boldsymbol{\varphi}, \boldsymbol{x}; \boldsymbol{\theta}^{(k)}\right)\right)^2\right], \tag{7}$$

where the target optimal Q-function R is given by

$$R = u^{(k)} + \gamma \max_{\boldsymbol{x}' \in \mathcal{X}} Q\left(\boldsymbol{\varphi}', \boldsymbol{x}'; \boldsymbol{\theta}^{(k-1)}\right), \tag{8}$$

and $\boldsymbol{\varphi}'$ is the next state sequence.

The gradient of the loss function with respect to the weights $\boldsymbol{\theta}^{(k)}$ is given by

$$\begin{aligned} \nabla_{\boldsymbol{\theta}^{(k)}} L\left(\boldsymbol{\theta}^{(k)}\right) &= \mathbb{E}_{\boldsymbol{\varphi}, \boldsymbol{x}, u, \boldsymbol{\varphi}'}\left[R \nabla_{\boldsymbol{\theta}^{(k)}} Q\left(\boldsymbol{\varphi}, \boldsymbol{x}; \boldsymbol{\theta}^{(k)}\right)\right] \\ &- \mathbb{E}_{\boldsymbol{\varphi}, \boldsymbol{x}}\left[Q\left(\boldsymbol{\varphi}, \boldsymbol{x}; \boldsymbol{\theta}^{(k)}\right) \nabla_{\boldsymbol{\theta}^{(k)}} Q\left(\boldsymbol{\varphi}, \boldsymbol{x}; \boldsymbol{\theta}^{(k)}\right)\right]. \end{aligned} \tag{9}$$

This process repeats B times and $\boldsymbol{\theta}^{(k)}$ is then updated according to these randomly selected experiences.

V. Performance analysis

We prove the convergence of the proposed two-dimensional anti-jamming scheme to the optimal strategy in the dynamic game and provide a performance bound of the utility of the mobile device against jamming attack. For simplicity, the channel gain between the jammers and the new radio node is assumed to be h_J' and $\varrho = \sigma + P_f \eta$. The SINR is assumed to follow

$$\text{SINR}^{(k)} = \frac{P_s^{(k)} h_{s,f}^{(k)} \lambda^{(k)}}{\sigma + P_f \eta^{(k)} + \sum_{j=1}^{J} P_J^{(k)} h_{j,y_j}^{(k)} \mathcal{F}\left(f^{(k)} - y_j^{(k)}\right)}. \tag{10}$$

Theorem 1. *The fast-DQN based mobile communication scheme in Algorithm 1 achieves the optimal anti-jamming communication strategy and the performance is given by*

$$\boldsymbol{x}^* = [P, 1], \tag{11}$$

$$u = \frac{Ph_s\lambda}{N(\varrho + P_J h'_J)} + \frac{(N-1)Ph_s\lambda}{N\varrho} - C_p P - C_m, \tag{12}$$

if the jammer in the dynamic game randomly chooses its jamming channel, and if

$$C_m \leq \frac{P_s h_s \lambda P_J (h_J - h'_J)}{N(\varrho + P_J h'_J)(\varrho + P_J h_J)} \tag{13}$$

$$C_p \leq \frac{h_s\lambda(N\varrho + (N-1)P_J h_J)}{N\varrho(\varrho + P_J h_J)}. \tag{14}$$

Proof: By (10), if (14) holds, we have

$$\begin{aligned} u(\phi = 0) &= \frac{P_s h_s \lambda}{\varrho + P_J h_J \mathcal{F}(f - y_j)} - C_p P_s \\ &= \frac{P_s h_s \lambda}{N(\varrho + P_J h_J)} + \frac{(N-1) P_s h_s \lambda}{N\varrho} - C_p P_s \\ &\leq \frac{P_s h_s \lambda}{\varrho + P_J h'_J \mathcal{F}(f - y_j)} - C_p P_s - C_m \\ &= \frac{P_s h_s \lambda}{N(\varrho + P_J h'_J)} + \frac{(N-1) P_s h_s \lambda}{N\varrho} - C_p P_s - C_m \\ &= u(\phi = 1). \end{aligned}$$

If (14) holds, we have

$$\begin{aligned} \frac{\partial u}{\partial P_s} &= \frac{h_s\lambda}{\varrho + P_J h_J \mathcal{F}(f - y)} - C_p \\ &= \frac{h_s\lambda}{N(\varrho + P_J h_J)} + \frac{(N-1) h_s\lambda}{N\varrho} - C_p \\ &= \frac{h_s\lambda(N\varrho + (N-1)P_J h_J)}{N\varrho(\varrho + P_J h_J)} - C_p \geq 0. \end{aligned} \tag{15}$$

Therefore, we have $\arg\max_{\boldsymbol{x}} u = [P, 1]$, and by (10), we have (12). ■

Remark 1 If the mobile device has good channel conditions and a large number of the frequency channels with low transmit cost C_p as shown in (14), the utility of the mobile device linearly increases with P_s as shown in (15) and the mobile device uses the maximal transmit power P. If the jammer cannot block the backup radio node and the mobility cost C_m is low as shown in (13), the mobile device will move to a new location with $\phi = 1$ to maximize its utility given by (12).

Theorem 2. *The fast-DQN based mobile communication scheme in Algorithm 1 achieves the optimal anti-jamming communication strategy and the performance is given by*

$$\boldsymbol{x}^* = [P, 1], \tag{16}$$

$$u = \frac{Ph_s\lambda}{N^2} Z_2 - C_p P - C_m, \tag{17}$$

if a jammer randomly chooses its jamming channel and another sweep jammer blocks N_J neighboring channels in the dynamic game, and if

$$C_m \leq \frac{P_s h_s \lambda P_J}{N^2 Z_1}(h_J - h'_J) \tag{18}$$

$$C_p \leq \frac{h_s\lambda}{N^2} Z_2, \tag{19}$$

where

$$\begin{aligned} Z_1 = {} & \frac{N - N_J}{(\varrho + P_J h'_J)(\varrho + P_J h_J)} \\ & + \frac{N_J^2(N-1)}{(N_J\varrho + P_J h'_J)(N_J\varrho + P_J h_J)} \\ & + \frac{N_J^2(N_J + 1)}{(N_J\varrho + P_J h'_J(N_J+1))(N_J\varrho + P_J h_J(N_J+1))} \end{aligned}$$

$$\begin{aligned} Z_2 = {} & \frac{(N-1)(N-N_J)}{\varrho} + \frac{N - N_J}{\varrho + P_J h_J} \\ & + \frac{(N-1)N_J^2}{N_J\varrho + P_J h_J} + \frac{N_J^2}{N_J\varrho + P_J h_J(N_J+1)}. \end{aligned}$$

Proof: By (10), if (18) holds, we have

$$\begin{aligned} u(\phi = 0) = {} & \frac{P_s h_s \lambda}{N^2}\left(\frac{N - N_J}{\varrho + P_J h_J} + \frac{(N-1)N_J^2}{N_J\varrho + P_J h_J}\right. \\ & \left. + \frac{N_J^2}{N_J\varrho + P_J h_J(N_J+1)} + \frac{(N-1)(N-N_J)}{\varrho}\right) \\ & - C_p P_s \leq \frac{P_s h_s \lambda}{N^2}\left(\frac{N - N_J}{\varrho + P_J h'_J} + \frac{(N-1)N_J^2}{N_J\varrho + P_J h'_J}\right. \\ & \left. + \frac{(N-1)(N-N_J)}{\varrho} + \frac{N_J^2}{N_J\varrho + P_J h'_J(N_J+1)}\right) \\ & - C_p P_s - C_m = u(\phi = 1). \end{aligned}$$

If (19) holds, we have

$$\begin{aligned} \frac{\partial u}{\partial P_s} = {} & \frac{h_s\lambda}{\varrho + P_J h_J \mathcal{F}(f - y)} - C_p \\ = {} & \frac{(N - N_J)h_s\lambda}{N^2(\varrho + P_J h_J)} + \frac{(N-1)N_J^2 h_s\lambda}{N^2(N_J\varrho + P_J h_J)} \\ & + \frac{N_J^2 h_s\lambda}{N^2(N_J\varrho + P_J h_J(N_J+1))} + \frac{(N-1)(N-N_J)h_s\lambda}{N^2\varrho} \\ & - C_p = \frac{h_s\lambda}{N^2} Z_2 - C_p \geq 0. \end{aligned} \tag{20}$$

Therefore, we have $\arg\max_{\boldsymbol{x}} u = [P, 1]$, and by (10), we have (17). ■

Remark 2 If the mobile device has good channel conditions and a large number of the frequency channels with low transmit cost C_p as shown in (19), the utility of the mobile device linearly increases with P_s as shown in (20) and the mobile device uses the maximal transmit power P. If the jammer cannot block the backup radio node and the mobility cost C_m is low as shown in (18), the mobile device will move to a new location with $\phi = 1$ to maximize its utility given by (17).

The complexity of this fast-DQN based mobile communication scheme in Algorithm 1 denoted by Γ is quadratic in both the filter size and the number of the filters of the CNN. Let F_{l-1} be the number of the input channels of the CNN in

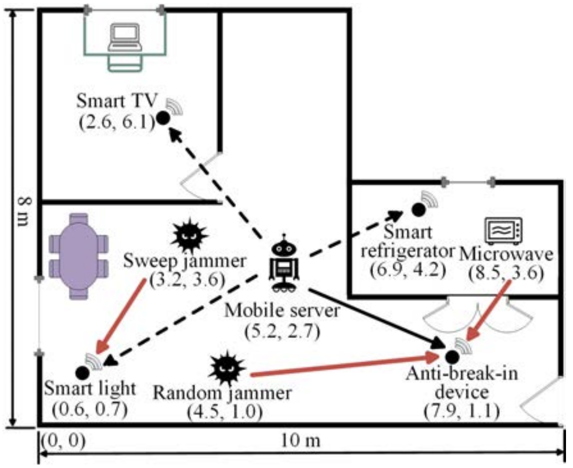


Fig. 3: Simulation topology in the command dissemination of a mobile server against a random jammer, a sweep jammer and an interference source.

Algorithm 1, F_l be the number of filters, N_l be the spatial size of the filter of Conv layer l and M_l be the output size of Conv layer l.

Theorem 3. *The computational complexity of the fast-DQN based mobile communication scheme in Algorithm 1 is given by*

$$\begin{aligned}\Gamma =&\mathcal{O}\Bigg(N_1^2F_1\left(\frac{N_C-N_1}{n_1}+1\right)^2\\&+F_1N_2^2F_2\left(\frac{N_C-N_1}{n_1n_2}-\frac{N_2-1}{n_2}+1\right)^2\Bigg).\end{aligned} \quad (21)$$

Proof: According to [30], the total complexity of the fast-DQN based mobile communication scheme in Algorithm 1 is $\mathcal{O}\left(\sum_{l=1}^{2}F_{l-1}N_l^2F_lM_l^2\right)$. The first Conv layer includes F_1 filters each of size $N_1\times N_1$, stride n_1, an $N_C\times N_C$ matrix as the input, and F_1 feature maps as the output. The second Conv layer consists of F_2 filters each of size $N_2\times N_2$, stride n_2, and F_2 feature maps as the output. According to [31], the output size of the first Conv layer is $(N_C-N_1)/n_1+1$ and that of the second Conv layer is $(N_C-N_1)/(n_1n_2)-(N_2-1)/n_2+1$. Therefore, the complexity of this scheme is given by (21). ■

VI. Applications

The RL based 2-D mobile communication scheme can be implemented in different mobile networks to resist jamming attacks. We present three examples and show the simulation results as follows.

A. Command dissemination of a mobile server

The 2-D mobile communication scheme can be applied in the command dissemination of a mobile device in an apartment to smart devices such as an anti-break-in device at the door, a smart TV and a smart refrigerator. The mobile server chooses the communication policy in each time slot and moves in the apartment to send command messages to a device against jamming.

Static jammers can neither block the radio node at the new location nor block the feedback channel. On the other hand, even if a smart jammer blocks the feedback channel, the mobile device will move to a new location and connect with a new radio node due to the ϵ-greedy policy in Algorithm 1. More specifically, the communication between the mobile device in the new location and the new AP cannot be blocked by the static jammers staying in the previous location due to the large path-loss fading.

We conducted a simulation to verify the performance of our scheme against a random jammer fixed at (4.5, 1.0) m, a sweep jammer fixed at (3.2, 3.6) m and an interference source fixed at (8.5, 3.6) m as shown in Fig. 3. More specifically, random jammers selected the same jamming channel with a probability 0.9 and a new channel with a probability 0.1. Sweep jammers blocked $N_J=4$ channels simultaneously in each time slot, i. e., the jamming power on each channel is P_J/N_J. A microwave in the kitchen sent interference signals during the transmission of the mobile server with a probability 0.05. The channel power gain $\mathbf{h}_s$ changed from 0.1 to 0.8 every 500 time slots with each time slot lasting 10.08 ms. The primary user randomly used a channel in each time slot.

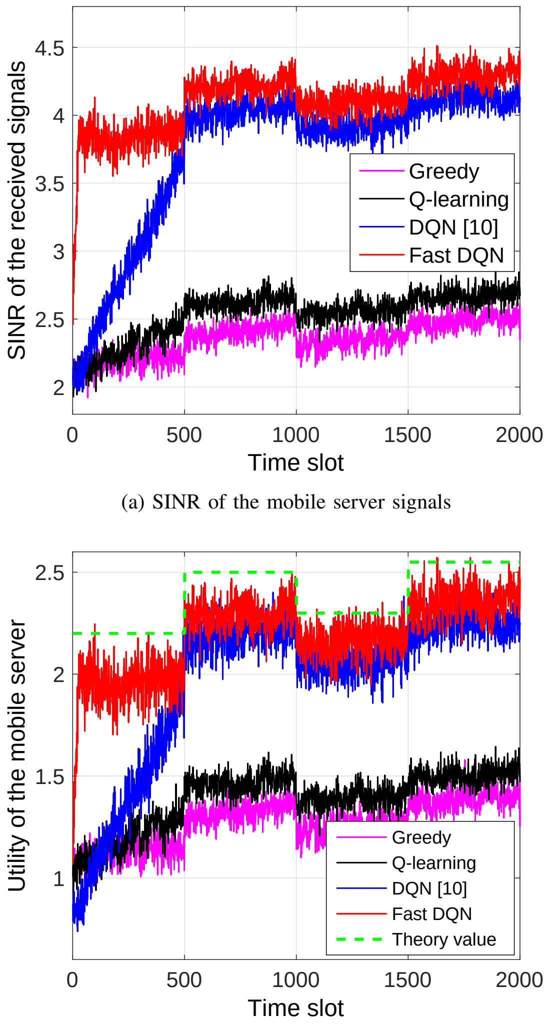
The mobile server was equipped with Intel i5-6200U CPU, 4GB RAM, and Ubuntu 14.04 64-bits system. In the simulations, $\sigma=1$, $C_m=0.8$, $C_p=0.2$, $C_h=0.4$, $\mathbf{h}_s^{(k)}\in[0,1]$,

Algorithm 2: Q-learning based 2-D mobile communication

```
Initialize $\gamma$, $\alpha$, $P_s$, $\mathrm{SINR}^{(0)}$, $\psi^{(1)}$, $\mathbf{Q}=\mathbf{0}$, and $\mathbf{V}=\mathbf{0}$
$\mathbf{s}^{(1)}=\left[\mathrm{SINR}^{(0)},\psi^{(1)}\right]$
for $k=1,2,\cdots$ do
    Choose $\boldsymbol{x}^{(k)}=\left[P_s^{(k)},\phi^{(k)}\right]$ via (4)
    if $\phi^{(k)}$ == 1 then
        Change the location and connect to a new radio node
    Observe the absence of PU and set $\lambda^{(k)}$
    if $\lambda^{(k)}$ == 0 then
        Keep silence
    else
        $f^{(k)}\leftarrow c_{\psi}^{k \bmod \kappa+1}$
        Send signals on channel $f^{(k)}$ with power $P_s^{(k)}$
    end
    Receive the $\mathrm{SINR}^{(k)}$ and $\psi^{(k+1)}$ on the feedback channel
    Obtain $u^{(k)}$ and $\mathbf{s}^{(k+1)}=\left[\mathrm{SINR}^{(k)},\psi^{(k+1)}\right]$
    Update $Q\left(\mathbf{s}^{(k)},\boldsymbol{x}^{(k)}\right)$ via (22)
    Update $V\left(\mathbf{s}^{(k)}\right)$ via (23)
end
```

TABLE II: CNN parameters in the mobile communication scheme in Algorithm 1

Layer	**Conv 1**	**Conv 2**	**FC 1**	**FC 2**
Input	$1*6*6$	$20*4*4$	320	128
Filter size	$3*3$	$2*2$	/	/
Stride	1	1	/	/
No. of filters	20	40	128	$2(L+1)$
Activation	ReLU	ReLU	ReLU	/
Output	$20*4*4$	$40*3*3$	128	$2(L+1)$



(a) SINR of the mobile server signals

(b) Utility of the mobile server

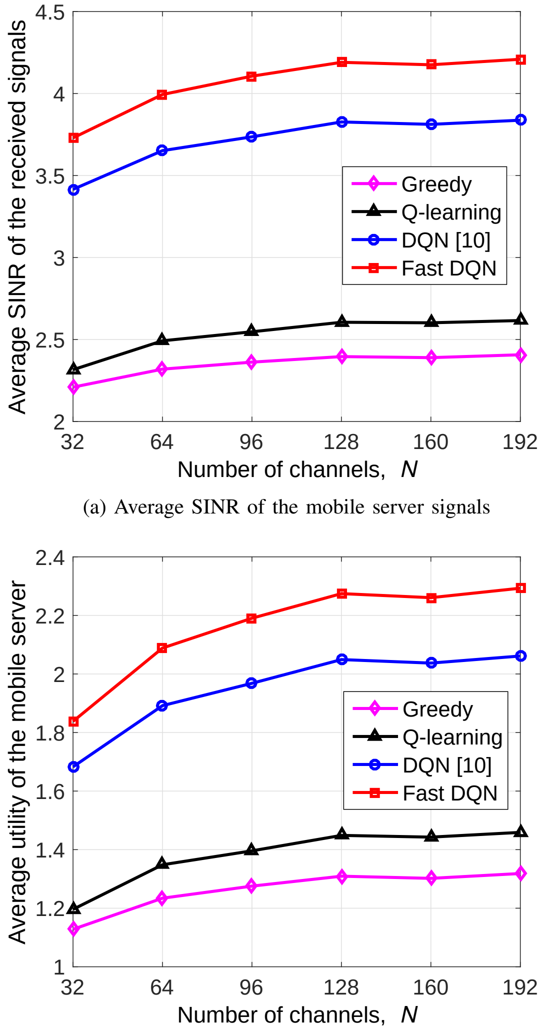
Fig. 4: Performance of the anti-jamming communication scheme in the commands dissemination of a mobile server with 96 frequency channels in a dynamic game against a random jammer, a sweep jammer and an interference source with $C_p = 0.2$ in the apartment as shown in Fig. 3.

$\mathbf{h}_j^{(k)} \in [0,1]$, $T = 300$, $N_r = 8$, $N_j = 4$, $L = 16$, $P = 8$, $P_j = 8$, $\kappa = 30$, $I = 200$, $K = 200$, $\vartheta = 10$, $\Phi = 4$ and $\zeta = 5$, if not specified otherwise. We set $W = 11$ to improve the communication efficiency and save the DQN memory overhead. According to the hyper parameters setting in [10], we set the minibatch size $B = 32$, ϵ linearly annealed from 0.5 to 0.05, and the discount factor γ linearly increased from 0.5 to 0.7 during the first 300 time slots for exploitation and was 0.7 afterwards. The CNN parameters were chose according to [10] as summarized in Table II.

As a benchmark, a Q-learning based 2-D anti-jamming mobile communication scheme as summarized in Algorithm 2 updates the Q-function according to the iterative Bellman equation as follows:

$$Q(\mathbf{s}, \boldsymbol{x}) \leftarrow \alpha\big(u + \gamma V\left(\mathbf{s}'\right)\big) + (1-\alpha)Q(\mathbf{s}, \boldsymbol{x}) \qquad (22)$$

$$V\left(\mathbf{s}\right) \leftarrow \max_{\boldsymbol{x}' \in \mathcal{X}} Q\big(\mathbf{s}, \boldsymbol{x}'\big), \qquad (23)$$



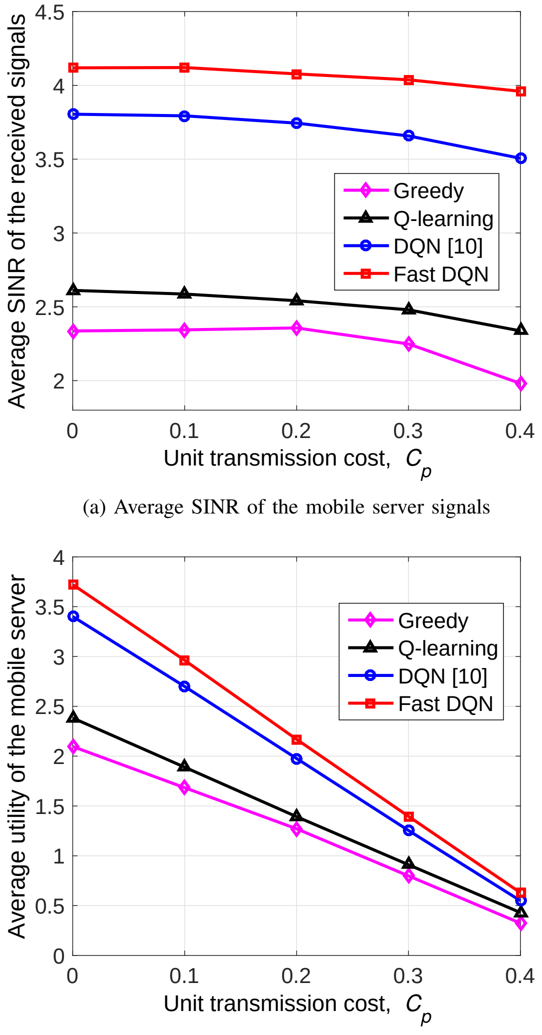
(a) Average SINR of the mobile server signals

(b) Average utility of the mobile server

Fig. 5: Average performance of the anti-jamming communication scheme in the commands dissemination of a mobile server with N frequency channels over 2000 time slots in each dynamic game and 200 scenarios against a random jammer, a sweep jammer and an interference source with $C_p = 0.2$ in the apartment as shown in Fig. 3.

where α is the learning rate that represents the weight of the current Q-function. Applying simulated annealing techniques similar to [10], the learning rate α in the Q-learning based scheme was linearly annealed from 0.7 to 0.5 during the first 300 time slots of the communication process in the simulations. Similarly, the discount factor γ linearly increased from 0.5 to 0.7 during the first 300 time slots of the communications for exploitation and was fixed at 0.7 afterwards.

As shown in Fig. 4, the fast-DQN based scheme achieves the performance given by Theorem 2 and outperforms other schemes with a higher SINR of the signals and a higher utility due to a faster learning speed. For instance, the fast DQN based scheme increases the SINR of the signals by 31.9% compared with the DQN based scheme, which is 76.2% and 84.7% higher than that of the Q-learning based and the greedy based schemes at 300-th time slot, respectively.



(a) Average SINR of the mobile server signals

(b) Average utility of the mobile server

Fig. 6: Average performance of the anti-jamming communication scheme in the commands dissemination of a mobile server with 96 frequency channels over 2000 time slots in each dynamic game and 200 scenarios against a random jammer, a sweep jammer and an interference source in the apartment as shown in Fig. 3.

Consequently, as shown in Fig. 4(b), the fast DQN based scheme improves the utility by 42.4%, 80.8% and 92.1% compared with the DQN based, the Q-learning based and the greedy based schemes at that time slot, respectively.

The anti-jamming performance of the proposed scheme improves with the number of channels as shown in Fig. 5. For example, the average SINR of the signals and the average utility of the mobile server are increased by the DQN based scheme by 12.1% and 21.8%, respectively, if the number of channels increases from 32 to 128. In addition, the DQN based scheme has much better performance than the Q-learning based and the greedy based schemes and the fast DQN based scheme can further improve the performance compared with the DQN based scheme. For instance, the DQN based scheme achieves 46.7% higher SINR and 41.0% higher utility compared with the Q-learning based scheme

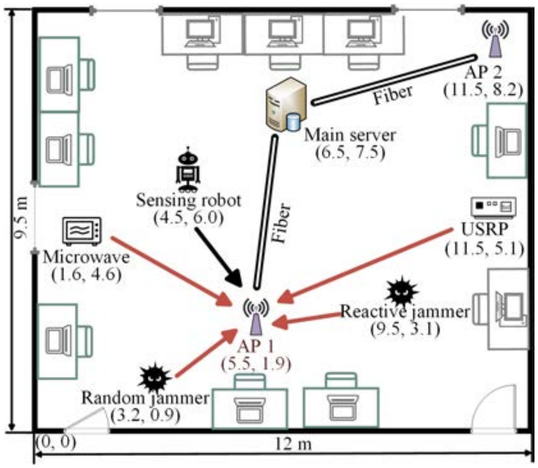
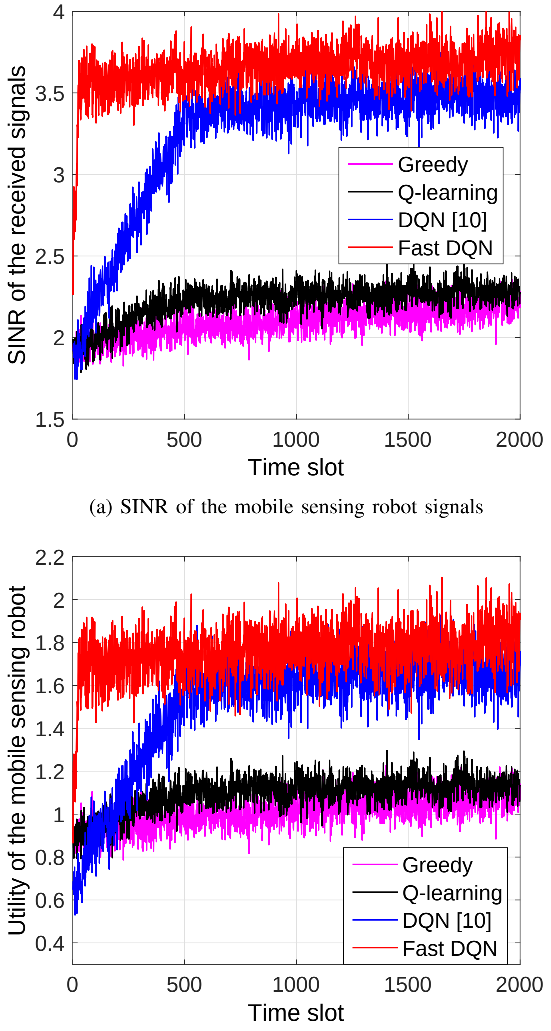


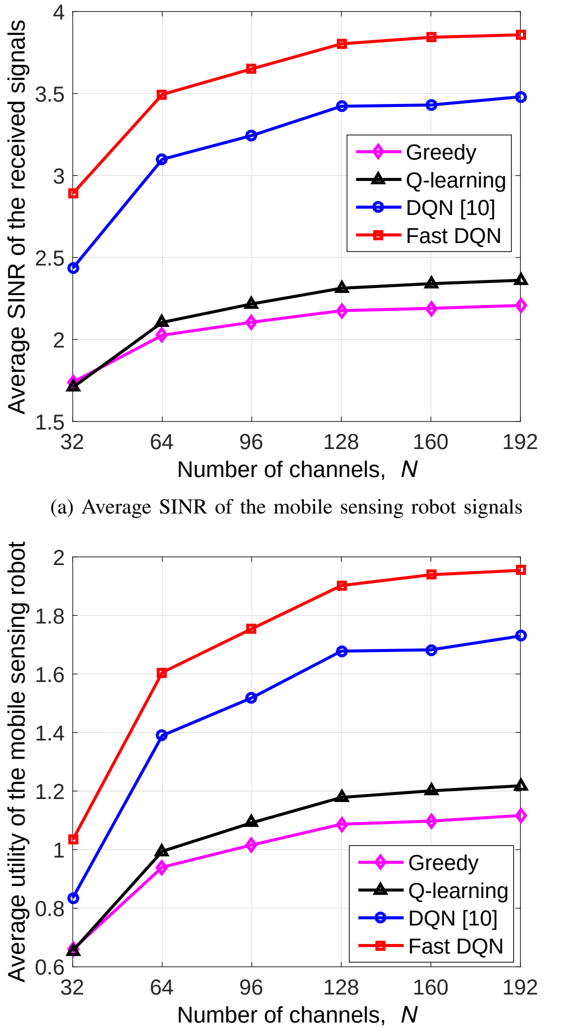
Fig. 7: Simulation topology in the sensing report collection of a sensing robot with two APs against a random jammer and a reactive jammer.



(a) SINR of the mobile sensing robot signals

(b) Utility of the mobile sensing robot

Fig. 8: Performance of the anti-jamming communication scheme in the sensing report transmission of a mobile sensing robot with 96 frequency channels in a dynamic game against a random jammer, a reactive jammer and two interference sources in the office as shown in Fig. 7.



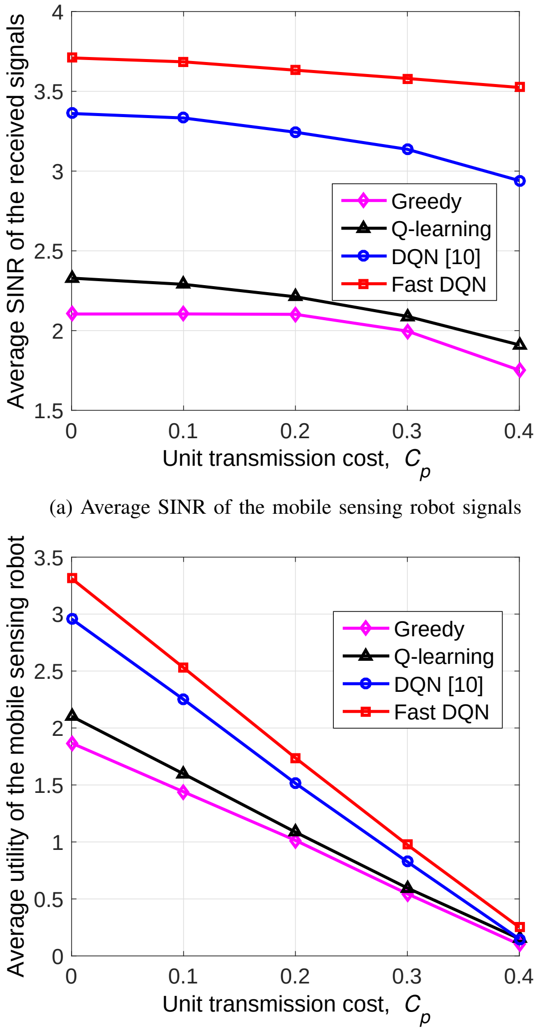
(a) Average SINR of the mobile sensing robot signals

(b) Average utility of the mobile sensing robot

Fig. 9: Average performance of the anti-jamming communication scheme in the sensing report transmission of a mobile sensing robot with N frequency channels over 2000 time slots in each dynamic game and 200 scenarios against a random jammer, a reactive jammer and two interference sources in the office as shown in Fig. 7.

for the system with 96 channels. Furthermore, the fast DQN based scheme increases the SINR of the signals by 73.8% and increases 71.7% utility, compared with the greedy based scheme for the system with 96 channels. On the other hand, the communication efficiency of the RL based communication scheme has to address the curse of the high-dimensionality under a large number of channels. For instance, the SINR and the utility of all the RL based schemes no longer improve with N if $N > 128$ as shown Fig. 5.

As shown in Fig. 6, both the SINR of the signals and the utility of the mobile server decrease with the unit transmission cost. For instance, the SINR of the signals and the utility of the mobile server decrease by the DQN based scheme by 3.9% and 63.1%, respectively, for the system with $C_p = 0.3$ instead of $C_p = 0$. In addition, the fast DQN based strategy always significantly outperforms other three schemes with different C_p. For instance, the fast DQN based scheme increases the SINR of the signals by 75.8% compared with the greedy based scheme, which is 59.3% and 8.6% higher than that of the Q-learning based and the DQN based schemes with $C_p = 0.1$, respectively. The fast DQN based scheme achieves 76.1%, 56.8% and 9.7% higher utility compared with the greedy based, the Q-learning based and the DQN based schemes, respectively.



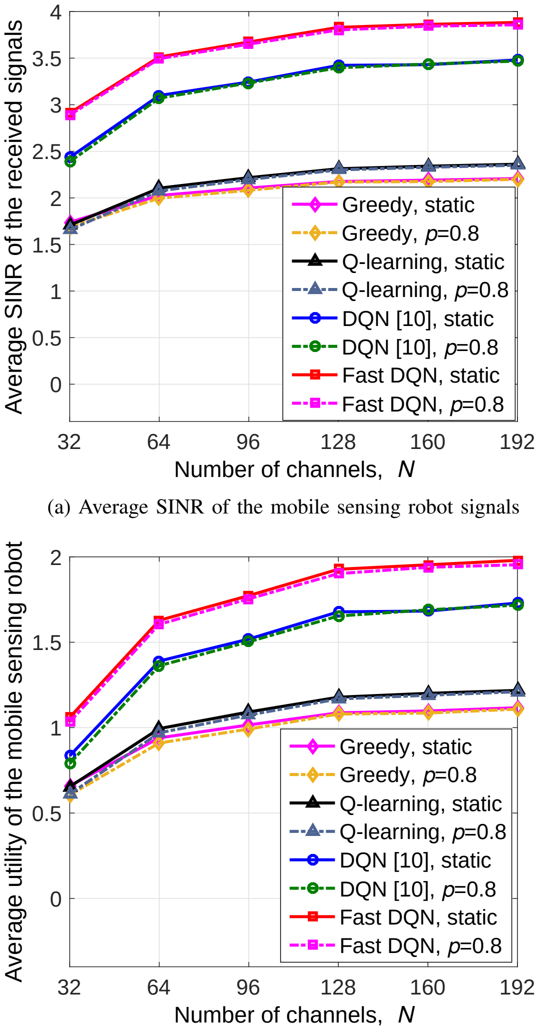
(a) Average SINR of the mobile sensing robot signals

(b) Average utility of the mobile sensing robot

Fig. 10: Average performance of the anti-jamming communication scheme in the sensing report transmission of a mobile sensing robot with 96 frequency channels over 2000 time slots in each dynamic game and 200 scenarios against a random jammer, a reactive jammer and two interference sources in the office as shown in Fig. 7.

In the simulation, the mobile device takes on average 2ms to update CNN weight parameters and choose the communication strategy. The data size is 100KB and the signal rate is 100Mb/s, the average transmission latency is 8ms, if the feedback time is 0.08ms and the feedback data size is 1KB.

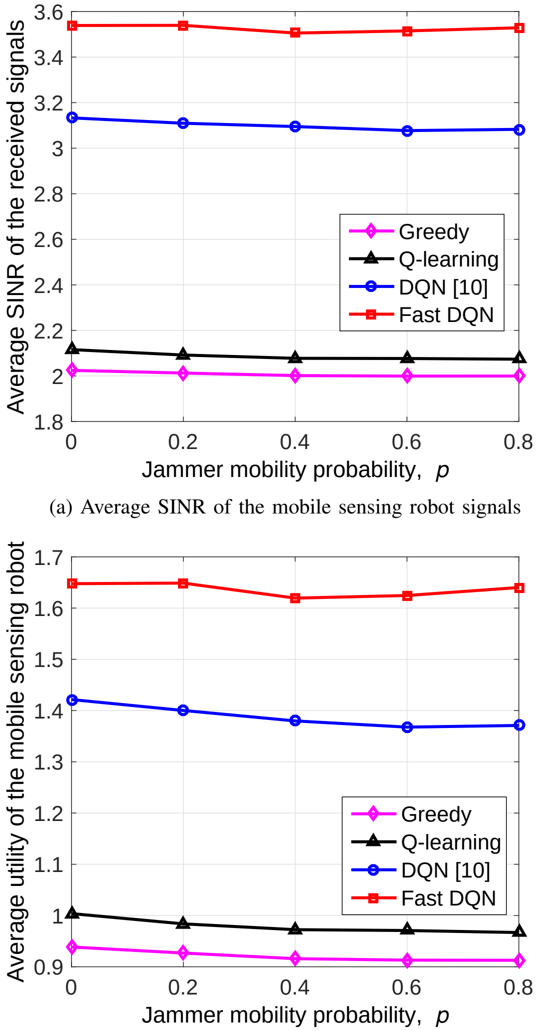
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2856854, IEEE Transactions on Vehicular Technology



(a) Average SINR of the mobile sensing robot signals

(b) Average utility of the mobile sensing robot

Fig. 11: Average performance of the anti-jamming communication scheme in the sensing report transmission of a mobile sensing robot with N frequency channels over 2000 time slots in each dynamic game and 200 scenarios against two mobile jammers and two interference sources with $p = 0.8$, in the office as shown in Fig. 7.



(a) Average SINR of the mobile sensing robot signals

(b) Average utility of the mobile sensing robot

Fig. 12: Average performance of the anti-jamming communication scheme in the sensing report transmission of a mobile sensing robot with 64 frequency channels over 2000 time slots in each dynamic game and 200 scenarios against two mobile jammers and two interference sources, in the office as shown in Fig. 7.

B. Sensing report collection

In the second application, a mobile sensing robot moves in the office to monitor the office and sends the sensing data over one of the N channels to the main server via two APs against jammers and interference sources. As shown in Fig. 7, a random jammer, a reactive jammer, a microwave and a universal software radio peripherals system were fixed at (3.2, 0.9) m, (9.5, 3.1) m, (1.6, 4.6) m and (11.5, 5.1) m, respectively. The reactive jammer continuously monitored $N_r = 8$ channels. The microwave interfered with the serving AP with a probability 0.1 and the USRP system interfered with the serving AP with a probability 0.05.

As shown in Fig. 8, the 2-D anti-jamming communication with the fast DQN based scheme outperforms the DQN based, the Q-learning based and the greedy based schemes, with a faster learning speed, a higher SINR of the signals, and a higher utility. For instance, the fast DQN based scheme converges after 50 time slots, which saves 90% and 99.999% of the learning time compared with the DQN based and the Q-learning based schemes, respectively. Therefore, the fast DQN based scheme increases the SINR of the signals by 24.1% compared with the DQN based scheme, which is 68.9% higher than that of the Q-learning based scheme at 300-th time slot. Consequently, as shown in Fig. 8(b), the fast DQN based scheme reaches the utility as high as 1.75 which is 39.7% and 78.9% higher than that of the DQN based and the Q-learning based schemes, respectively.

Fig. 9 shows that the proposed 2-D anti-jamming communication schemes can achieve higher SINR of the signals and higher utility of the mobile sensing robot with the number of channels increasing. For example, the average SINR of the signals with the fast DQN based scheme increases by 31.8% to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2018.2856854, IEEE Transactions on Vehicular Technology

3.81, and achieves 84.1% higher average utility, if the number of channels increases from 32 to 128. The utility of the fast DQN based scheme increases by 55.3% if the the number of channels increases from 32 to 64, and increases by 1.9% if the the number of channels increases from 128 to 160. In addition, the fast DQN based scheme has the highest average SINR of the signals and the highest average utility in all of the four schemes. For instance, the fast-DQN based scheme achieves 12.8% higher SINR of the signals compared with the DQN based scheme, which is 72.5% higher than that of the greedy based scheme for the system with 64 channels. Consequently, as shown in Fig. 9(b), the average utility of the mobile sensing robot with the fast DQN based scheme increases by 15.5% and 70.7% compared with the DQN based and the greedy based schemes, respectively.

Fig. 10 illustrates the impacts of the unit transmission cost on the performance showing that both the average SINR of the signals and the average utility of the robot decreases with the unit transmission cost. For instance, the DQN based scheme decreases the SINR of the signals by 4.9% and achieves 63.3% lower utility, if C_p increases from 0.1 to 0.3. In addition, the anti-jamming performance of the DQN based scheme exceeds that of the Q-learning based and the greedy based schemes, and can be further improved by the fast DQN based scheme. For example, the DQN based scheme achieves 58.4% higher SINR of the signals and 56.3% higher utility than that of the greedy based scheme, and be further increased by 16.7% and 19.4% with the fast DQN based scheme, for the system with $C_p = 0.1$.

C. Sensing report collection against mobile jammers

As shown in Fig. 7, two mobile jammers changed their locations randomly with a probability 0.8 every 200 time slots. The channel gains with the mobile jammers randomly changed with a probability 0.8 ranging from 0.28 to 0.9 every 200 time slots. As shown Fig. 11, the proposed schemes are robust against the mobile jammers. For instance, the average SINR and the utility of the robot with the fast DQN based scheme decrease by 0.6% and 1.1% if $N = 96$ compared with the static jammers.

Fig. 12 illustrates the impacts of the jamming mobility, showing that the proposed schemes are robust against jamming mobility. For example, the SINR of the signals of the fast DQN based scheme slightly decreases by 0.7% if the jammer mobility probability p increases from 0 to 0.6 as shown in Fig. 12(a). Consequently, as shown in Fig. 12(b), the utility of the robot slightly decreases by 1.4% if p increases from 0 to 0.6.

VII. Conclusions

In this paper, we have proposed an RL based frequency-space anti-jamming mobile communication system that exploits spread spectrum and user mobility to resist cooperative jamming and strong interference. We have shown that, by applying a DQN based frequency-space anti-jamming mobile communication scheme, a mobile device can achieve an optimal power allocation and moving policy, without being aware of the jamming and interference model and the radio channel model. Moreover, we have seen that the proposed fast DQN based 2-D mobile communication scheme combining hotbooting, DQN and macro-actions can further accelerate learning and thus improve the jamming resistance. Simulation results show that the fast DQN based scheme increases the SINR of the signals compared with the benchmark scheme [10]. For instance, the fast DQN based scheme saves 90% of the learning time required by DQN, and increases the SINR of the signals and the utility of the mobile device by 31.9% and 42.4%, respectively, compared with the DQN based scheme.

References

[1] A. Benslimane and H. Nguyen-Minh, “Jamming attack model and detection method for beacons under multichannel operation in vehicular networks,” *IEEE Trans. Vehicular Technology*, vol. 66, no. 7, pp. 6475–6488, July 2017.

[2] F. Zhu, F. Gao, M. Yao, and H. Zou, “Joint information- and jamming-beamforming for physical layer security with full duplex base station,” *IEEE Trans. Signal Processing*, vol. 62, no. 24, pp. 6391–6401, Dec. 2014.

[3] L. Xiao, C. Xie, M. Min, and W. Zhuang, “User-centric view of unmanned aerial vehicle transmission against smart attacks,” *IEEE Trans. Vehicular Technology*, vol. 67, no. 4, pp. 3420–3430, April 2018.

[4] Q. Wang, T. P. Nguyen, K. Pham, and H. M. Kwon, “Mitigating jamming attack: A game theoretic perspective,” *IEEE Trans. Vehicular Technology*, DOI: 10.1109/TVT.2018.2810865.

[5] J. Dams, M. Hoefer, and T. Kesselheim, “Jamming-resistant learning in wireless networks,” *IEEE/ACM Trans. Networking*, vol. 24, no. 5, pp. 2809–2818, 2016.

[6] M. Labib, S. Ha, W. Saad, and J. H. Reed, “A Colonel Blotto game for anti-jamming in the Internet of Things,” in *Proc. IEEE Global Comm. Conf. (GLOBECOM)*, pp. 1–6, San Diego, CA, Dec. 2015.

[7] S. D’Oro, E. Ekici, and S. Palazzo, “Optimal power allocation and scheduling under jamming attacks,” *IEEE/ACM Trans. Networking*, vol. 25, no. 3, pp. 1310–1323, 2017.

[8] L. Zhang, Z. Guan, and T. Melodia, “United against the enemy: Anti-jamming based on cross-layer cooperation in wireless networks,” *IEEE Trans. Wireless Comm.*, vol. 15, no. 8, pp. 5733–5747, Aug. 2016.

[9] N. Adem and B. Hamdaoui, “Jamming resiliency and mobility management in cognitive communication networks,” in *Proc. IEEE Int’l Conf. on Communications (ICC)*, pp. 1–6, Paris, France, May 2017.

[10] G. Han, L. Xiao, and H. V. Poor, “Two-dimensional anti-jamming communication based on deep reinforcement learning,” in *Proc. IEEE Int’l Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 1–5, New Orleans, LA, Mar. 2017.

[11] A. S. Lakshminarayanan, S. Sharma, and B. Ravindran, “Dynamic action repetition for deep reinforcement learning.,” in *Proc. AAAI Conf. on Artificial Intelligence (AAAI)*, pp. 2133–2139, San Francisco, California, Feb. 2017.

[12] Y. Wu, B. Wang, K. J. R. Liu, and T. C. Clancy, “Anti-jamming games in multi-channel cognitive radio networks,” *IEEE Journal on Selected Areas in Comm.*, vol. 30, no. 1, pp. 4–15, Jan. 2012.

[13] L. Xiao, Y. Li, J. Liu, and Y. Zhao, “Power control with reinforcement learning in cooperative cognitive radio networks against jamming,” *Journal of Supercomputing*, vol. 71, no. 9, pp. 3237–3257, Apr. 2015.

[14] X. Tang, P. Ren, Y. Wang, Q. Du, and L. Sun, “Securing wireless transmission against reactive jamming: A Stackelberg game framework,” in *Proc. IEEE Global Comm. Conf. (GLOBECOM)*, pp. 1–6, San Diego, CA, Dec. 2015.

[15] L. Xiao, J. Liu, Q. Li, N. B. Mandayam, and H. V. Poor, “User-centric view of jamming games in cognitive radio networks,” *IEEE Trans. Information Forensics and Security*, vol. 10, no. 12, pp. 2578–2590, Dec. 2015.

[16] R. El-Bardan, V. Sharma, and P. K. Varshney, “Learning equilibria for power allocation games in cognitive radio networks with a jammer,” in *Proc. IEEE Global Conf. on Signal and Information Processing (GlobalSIP)*, pp. 1–6, Washington, DC, Dec. 2016.

[17] B. Wang, Y. Wu, K. J. R. Liu, and T. C. Clancy, “An anti-jamming stochastic game for cognitive radio networks,” *IEEE Journal on Selected Areas in Comm.*, vol. 29, no. 4, pp. 877–889, Mar. 2011.

[18] 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 and Information Processing over Networks*, vol. 2, no. 1, pp. 49–56, Mar. 2016.

[19] M. Hanawal, M. Abdelrahman, and M. Krunz, "Joint adaptation of frequency hopping and transmission rate for anti-jamming wireless systems," *IEEE Trans. Mobile Computing*, vol. 15, no. 9, pp. 2247–2259, Sep. 2016.

[20] C. Chen, M. Song, C. Xin, and J. Backens, "A game-theoretical anti-jamming scheme for cognitive radio networks," *IEEE Network*, vol. 27, no. 3, pp. 22–27, June 2013.

[21] Y. Gwon, S. Dastangoo, C. Fossa, and H. T. Kung, "Competing mobile network game: Embracing anti-jamming and jamming strategies with reinforcement learning," in *Proc. IEEE Conf. on Comm. and Network Security (CNS)*, pp. 28–36, National Harbor, MD, Oct. 2013.

[22] F. Slimeni, B. Scheers, Z. Chtourou, and V. L. Nir, "Jamming mitigation in cognitive radio networks using a modified Q-learning algorithm," in *Proc. IEEE Int'l Conf. on Military Communications and Information Systems*, pp. 1–7, Cracow, Poland, May 2015.

[23] T. Chen, J. Liu, L. Xiao, and L. Huang, "Anti-jamming transmissions with learning in heterogenous cognitive radio networks," in *Proc. IEEE Wireless Comm. and Networking Conference Workshops, So-HetNets Workshop*, pp. 293–298, New Orleans, LA, June 2015.

[24] S. Singh and A. Trivedi, "Anti-jamming in cognitive radio networks using reinforcement learning algorithms," in *Proc. IEEE Int'l Conf. on Wireless and Optical Comm. Networks (WOCN)*, pp. 1–5, Indore, India, Nov. 2012.

[25] B. F. Lo and I. F. Akyildiz, "Multiagent jamming-resilient control channel game for cognitive radio ad hoc networks," in *Proc. IEEE Int'l Conf. on Communications (ICC)*, pp. 1821–1826, Ottawa, Canada, Jun. 2012.

[26] M. A. Aref, S. K. Jayaweera, and S. Machuzak, "Multi-agent reinforcement learning based cognitive anti-jamming," in *Proc. IEEE Wireless Comm. and Networking Conf. (WCNC)*, pp. 1–6, San Francisco, CA, May 2017.

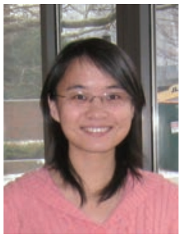
[27] X. He, H. Dai, and P. Ning, "Faster learning and adaptation in security games by exploiting information asymmetry," *IEEE Trans. Signal Processing*, vol. 64, no. 13, pp. 3429–3443, Jul. 2016.

[28] O. B. Akan, O. Karli, and O. Ergul, "Cognitive radio sensor networks," *IEEE Network*, vol. 23, no. 4, pp. 34–40, Aug. 2009.

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

[30] K. He and J. Sun, "Convolutional neural networks at constrained time cost," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 5353–5360, Boston, MA, June 2015.

[31] C. C. T. Mendes, V. Frémont, and D. F. Wolf, "Exploiting fully convolutional neural networks for fast road detection," in *Proc. IEEE Int'l Conf. on Robotics and Automation (ICRA)*, pp. 3174–3179, Stockholm, Sweden, May 2016.



Liang Xiao (M'09, SM'13) is currently a Professor in the Department of Communication Engineering, Xiamen University, Fujian, China. She has served in several editorial roles, including an associate editor of IEEE Trans. Information Forensics & Security and IET Communications. Her research interests include wireless security, smart grids, and wireless communications. She won the best paper award for 2016 IEEE INFOCOM Bigsecurity WS. She received the B.S. degree in communication engineering from Nanjing University of Posts and T-elecommunications, China, in 2000, the M.S. degree in electrical engineering from Tsinghua University, China, in 2003, and the Ph.D. degree in electrical engineering from Rutgers University, NJ, in 2009. She was a visiting professor with Princeton University, Virginia Tech, and University of Maryland, College Park. She is a senior member of the IEEE.

Donghua Jiang received the B.S. degree in electronic information science and technology from Xiamen University, Xiamen, China, in 2017, where she is currently pursuing the M.S. degree with the Department of Communication Engineering. Her research interests include network security and wireless communications.

Dongjin Xu received the B.S. degree in communication engineering from Xiamen University, Xiamen, China, in 2016, where she is currently pursuing the M.S. degree with the Department of Communication Engineering. Her research interests include network security and wireless communications.

Hongzi Zhu (M'07) received his Ph.D. degree in computer science from Shanghai Jiao Tong University in 2009. He was a Post-doctoral Fellow in the Department of Computer Science and Engineering at Hong Kong University of Science and Technology and the Department of Electrical and Computer Engineering at University of Waterloo in 2009 and 2010, respectively. He is now an associate professor at the Department of Computer Science and Engineering in Shanghai Jiao Tong University. His research interests include vehicular networks, network and mobile computing. He received the Best Paper Award from IEEE Globecom 2016. He is a member of the IEEE Computer Society and Communication Society.

Yanyong Zhang (M'08, SM'15, F'17) received her B.S. from the University of Science and Technology of China (USTC) in 1997, and Ph.D. from Penn State University in 2002. From 2002 and 2018, she was on the faculty of the Electrical and Computer Engineering Department at Rutgers University. She was also a member of the Wireless Information Networks Laboratory (Winlab). Since July 2018, she joined the school of Computer Science and Technology at USTC.

She has 21 years of research experience in the areas of sensor networks, ubiquitous computing, and high-performance computing, and has published more than 110 technical papers in these fields. She received the NSF CAREER award in 2006, and was elevated to IEEE Fellow in 2017. She currently serves as the Associate Editor for several journals, including IEEE/ACM Transactions on Networking, IEEE Transactions on Mobile Computing, IEEE Transactions on Service Computing, and Elsevier Smart Health.

H. Vincent Poor (S'72, M'77, SM'82, F'87) received the Ph.D. degree in EECS 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 currently 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 universities, including 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. He received the Marconi and Armstrong Awards of the IEEE Communications Society in 2007 and 2009, respectively. Recent recognition of his work includes the 2017 IEEE Alexander Graham Bell Medal, Honorary Professorships at Peking University and Tsinghua University, both conferred in 2017, and a D.Sc. *honoris causa* from Syracuse University also awarded in 2017.