Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Energy Efficient Transmission Power Control Policy of the Delay Tolerable Communication Service

RUI ZHU[1], JIANXIN GUO[1], FENG WANG[1], BAOQIN LIN[1] AND YANGCHAO HUANG[2].

[1]The School of Information Engineering, Xijing University, Xi'an 710123, China.

[2]The Institute of Information and Navigation, Air Force Engineering University, Xi'an 710008, China

Corresponding author: Feng Wang (e-mail: wangfengisn@163.com).

This work was supported by the Nature Science Foundation of Shaanxi Province (Grant No. 2018JM6075), Shaanxi Key Laboratory of Integrated and Intelligent Navigation open fund (Grant No.SKLIIN-20180211) and the research foundation for talented scholars of Xijing University (Grant No.XJ19B01 and No. XJ17B06), the Key Research and Development Plan Project of Shaanxi Provincial Science and Technology Department (No.2018ZDXM-NY-014) and Xi'an Science and Technology Project (No.201805043YD21CG27(1))

ABSTRACT In recent years, the development of wireless communication leads to an explosive growth of energy demand, widely application of smart devices and rapid emergence of services. So, the energy efficient communication is expected urgently to save power and prolong the lifetime of the resource-constrained terminal devices. Especially in 5G age, the excellent information transmission rate provides more opportunity to save power by adjusting the transmission power for the delay tolerable (DT) service. Meanwhile, although the tradeoff between energy efficiency and service delay plays a non-negligible role in the energy efficient communication, it is not exploited sufficiently due to the time variation and randomness of wireless communication channel. For this reason, the fundamental tradeoff between energy efficiency and delay of the DT service is investigated and analyzed. And, the optimal problem of energy efficient communication for DT service is formulated as a Markov Decision Process (MDP) problem which can be solved effectively by statistical dynamic programming (SDP) since the perfect channel state information (CSI) is hard to obtain. To improve the utility of research, the approximate SDP (ASDP) and Q-learning are also investigated to overcome the limitation of the curse of dimensionality and model-based algorithm respectively.

INDEX TERMS Energy efficient wireless communication, power allocation, statistic dynamic programming, approximate statistic dynamic programming, delay tolerable service, Q-learning

I. INTRODUCTION

WITH the fast development of the wireless communication and computer technologies, tens of billions smart devices, such as sensors and wearable devices, are enabled to interact with each other and provide various services for us. Meanwhile, this trend also leads to an explosive growth of energy demand and carbon emissions in future. The report of SMART2020 [1] showed that the CO2 levels in the atmosphere will increase sharply in recently. And, [2] estimated the amount of carbon dioxide (CO2) caused by cellular networks will reach 345 million tons in 2020. All of these carbon emissions will affect the environment acutely. Furthermore, the increasing of the power consumption also increases the operational cost of the mobile network operator. According to the conclusions of [3], almost 30 percentages of the Telecom operational cost was attributed to the energy consumption. In order to reduce the power consumption and operational cost of wireless communication, the energy efficient wireless communication, which also be known as green communication, was emerged.

Besides the ecological and economic factors, the battery life of terminal devices was another non-negligible reason for heading towards energy efficient communication. During the past years, the battery technology developed at a comparatively slower rate while the communication service grew rapidly. For this reason, many services, such as Internet of Things (IoT) and Device-to-Device (D2D), were fundamentally restricted by the devices with limited battery. Energy efficient communication provides an effective method to prolong the live life of power-constraint terminal devices.

Technologies towards energy-efficient wireless networks have been studied for a long term. In [4], Chen *et al.* discussed the tradeoff among energy efficiency, spectrum efficiency, convergence efficiency and service delay of cellular network. The conclusions showed that the increasing of energy efficiency, unlike the increasing of spectrum efficiency which is almost always beneficial for the Quality of Service (QoS), may lead to the degradation of QoS. In [5], the energy consumption model of macro cell Base Station (BS) was formulated based on the analysis of the fundamental circuit. Moreover, R.Mahapatra *et al.* simplify the energy consumption model of macro cell BS by dividing the consumed energy into static power and transmission power to reduce the complexity of analysis [6]. In [7], a serial indices of the energy efficiency are proposed in terms of network convergence, throughput and spectrum efficiency. With the maturity of the research frame of the energy efficient communication, researchers began to investigate the specific technology to realize the energy efficient communication.

In this paper, we investigated the potential of the energy efficiency for the Delay Torlerable (DT) service with partial Channel States Information (CSI). We also proposed three transmission power control policies under various assumption. Our main contributions are summarized as follows:

- The energy efficient problem of the DT service is formulated. Based on the formulated optimal problem and convex optimal theory, the performance upper bound of the transmission power is derived.
- Since the perfect CSI of the future cannot be obtained, the energy efficient problem of the DT is reformulated as a Markov Decision Problem (MDP) with statistical CSI. Furthermore, under the discrete channel assumption, the statistical dynamic programming (SDP) is adopted to design the energy efficient transmission policy for the proposed MDP.
- Due to the dimension curse, the SDP is not available for the continuous channel. Unfortunately, the channel model of wireless communication follows continuous distributions in most of the general case. For this reason, we adopted the approximate SDP (ASDP) to design the energy efficient transmission policy for the DT service under the continuous channel assumption.
- The policies of the ASDP and SDP are developed based on the statistical channel model, which may not be available in practice. In order to improve the utility value of research, we develop a model free energy efficient transmission policy based on Q-learning.

The remainder of this paper is organized as follows. In Section II, the related works of energy efficient communication are brief reviewed. In Section III, we formulate the system model of DT service. And the optimal problem of the transmission power is also proposed to derive the performance bound under the perfect CSI assumption. In Section IV, we reformulate the proposed optimal problem as a MDP since the perfect CSI is unavailable in practical scenarios. Furthermore, the SDP is adopted to solve the proposed MDP under the discrete channel assumption. In order to further increase the utility value of our research, the ASDP is adopted to deal with the dimension curse caused by the continuous channel in Section V. And, the Q-learning is investigated to overcome the limitation of model-based policy. Section VI draws a conclusion and summarizes the paper.

II. RELATED WORKS

The first popular topic of energy efficient communication is the optimal network planning because the number of BS sites almost approaches 11.2 million by 2020. All of these BS sites will draw the energy consumption significantly. In [8], the sleep mechanism was proposed to save the power by components deactivation when they are not working. Following this idea, [9] designed a sleep mode to save power by turning off the BS with less loading. During the sleep mode, the power consumption of the BS is very low. When the loading increases, the sleeping BS will be awoken to ensure the QoS. In order to optimize the mode switching policy of BS, S Zhou *et al.* developed an adjustment policy of BS modes based on the traffic intensity under the blocking probability constraint in [10] [11].With the explosive rise in the number of subscribers and traffic [12], the concept of cell breathing emerged as the generalized BS sleeping mode. The cell breathing can save power by adaptive adjustment of cell size in accordance with the loading and Qos variations. In [13], a cell breathing algorithm was proposed and evaluated. The results showed that the cell breathing is an effective method to enhance the energy efficiency of cellular network. In [14], the cell breathing was adopted by the green heterogeneous networks and showed great performance. In [15] [16], the liquid cell management techniques were developed based on the cell breathing to save power under various QoS constraints.

Besides the optimal network planning, the resource allocation was another effective method to improve the energy efficiency. The most well known power allocation algorithms were the equal power allocation (EPA) [17] [18] and the water filling (WF) algorithms [19]. In [20], the Double-threshold WF algorithm was proposed to improve the spectrum efficiency and energy efficiency in the spectrum sharing systems with statistical CSI. In [21], a multi-objective optimal problem was formulated to improve the energy efficiency and spectrum efficiency of the cognitive radio networks. In [22], a joint optimal problem of the OFDMA system is formulated. And, the researcher also proposed an iterative algorithm to solve the proposed optimal problem. However, the traditional WF algorithm and several of its derivation algorithms were designed to improve the system performance by optimizing the power allocation in frequency and space domain. To further investigate the potential of energy efficiency in the time domain, a few researchers studied the tradeoff between energy efficiency and service delay. In [23], a joint link scheduling and power allocation algorithm was

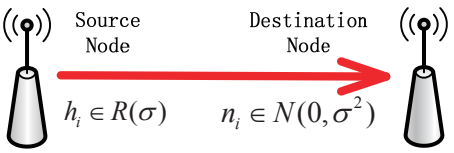


FIGURE 1. An illustration of the point to point communication link with a source node and a destination node.

proposed for the D2D-assisted wireless caching networks. In [24], a time-domain resource allocation scheme was proposed to improve the energy efficiency performance of D2D in the delay-insensitive scenario. Moreover, in [25] [26] [27], various energy efficient policies are studied based on the unlimited queue length model.

But to the best of our knowledge, the most existing researches of resource allocation over time-domain were developed under the simple wireless channel model or perfect CSI assumptions to reduce the research complexity. These assumptions do not correspond to the practical wireless communication scenarios. Under the general conditions, the obtained CSI is imperfect and partial since the wireless channel is random and time-varying. Fortunately, the development of reinforcement learning provides an effective tool to improve the energy efficiency of wireless communication with partial CSI [28] [29] [30]. And, a few researchers have employed it to optimize the energy efficiency of various wireless networks [31] [32]. Motivated by their work, the reinforcement learning was adopted in this paper to improve the energy efficiency of the DT service.

III. SYSTEM MODEL AND PROBLEM FORMULATION

In this section, the model of DT service is formulated. And, the transmission power optimal problem of DT service is also proposed.

A. SYSTEM MODEL

Just as the Fig. 1 shown, a point-to-point communication link with the multiplicative fading and the additive white Gaussian noise is studied in this paper. So, the receiving signal of the destination node can be presented as

$$y_i = \sqrt{P_{t,i}} h_i x_i + n_i \quad i = 1, 2, ... \tag{1}$$

where $P_{t,i}$ is the transmission power of i th slot. h_i and n_i are the channel fading coefficient and Gaussian noise variable, respectively. y_i and x_i are the receiving signal and transmission signal. The proposed link is supposed to provide a serials DT service, such as the video play or file transfer. Unlike the delay sensitive service, the DT service requires to transfer a certain amount information in a period rather than immediately. So, the energy efficiency of DT service can be improved by optimizing the transmission power in time domain. The Fig. 2 shows an example of the optimal power allocation for the DT service which requires to transfer I bits

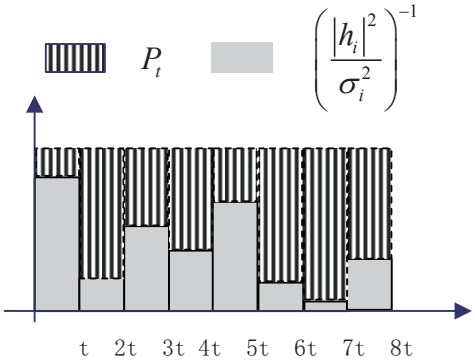


FIGURE 2. A power allocation pattern of a 8 slots delay constraint service.

in 8 slots. The solid bars in the Fig. 2 represent the channel condition reciprocal of the wireless channel which is defined as

$$\gamma_i^{-1} = \left(\frac{|h_i|^2}{\sigma_i^2}\right)^{-1}, \quad i = 1, 2, ...8. \tag{2}$$

Meanwhile, the line bars in the Fig. 2 represent the transmission power of the i th slot. This idea is similar to the well-known water filling algorithm except we allocate the transmission power resource in the time domain. In order to reduce the analysis complexity, we assume all of the nodes in the proposed model equip the single antenna.

B. FROM THE MAXIMAL ENERGY EFFICIENCY TO MINIMUM TRANSMISSION ENERGY

When the energy efficient communication is studied, a popular and precise indics is the energy efficiency which is defined as

$$\frac{C}{P_f + P_t} \quad (bit/Joule/Hz), \tag{3}$$

where C is the channel rate. P_f and P_t are the static power and transmission power respectively. The main job of the energy efficient communication is to maximize the energy efficiency under a set of constraint conditions. When we focus on a given DT service, the energy efficiency can be further represented as

$$\frac{I}{TP_f + \sum_{i=1}^{T} P_{t,i}} \quad (bit/joule/Hz), \tag{4}$$

where I and T are the required total information and the maximal tolerable delay respectively. And, $P_{t,i}$ is the transmission power in the i th slot as we defined before. Since the required information I and total static power TP_f are constants for a given service, the optimal problem of energy efficiency can be turned to the optimal problem of transmission power $\sum_{i=1}^{T} P_{t,i}$. So, in the following part, we will focus on the minimization problem of transmission power.

C. PROBLEM FORMULATED

Without loss of generality, the required information I of DT service can be represented as

$$I = \sum_{i=1}^{T} log_2(1 + P_{t,i}\gamma_i), \tag{5}$$

where γ_i is the channel condition which is defined in (2). And, the transmission power constraint is

$$P_{t,i} \leq P_{t,max}, \quad i = 1, 2, ..., T. \tag{6}$$

where $P_{t,max}$ is the available maximal transmission power of source node. So, the transmission power optimal problem of the DT service can be formulated as

$$\begin{aligned} min: \quad & \sum_{i-1}^{T} P_{t,i} \\ s.t.: \quad & \sum_{i=1}^{T} log_2(1 + P_{t,i}\gamma_i) \geq I. \\ & P_{t,i} \leq P_{t,max}, \quad i = 1, 2, ..., T. \end{aligned} \tag{7}$$

Without loss of generality, the sign '$\geq$' is adopted to replace the sign '=' in the information constraint without affecting the optimal result. Apparently, (7) is a classical convex optimal problem which can be solved effectively when the source node has the perfect CSI (γ) of the future T slots. However, it is almost impossible to obtain the perfect CSI of the future T slots in practical for the source node. So, (7) actually provides a transmission power lower bound or an energy efficiency upper bound of the practical DT service. With the development of various channel estimation, the source node can get a pretty well CSI of the current slot. This advantage means the source node should adjust the transmission power base on the observed CSI of the current slot. Based on this idea, we reformulated the transmission power optimal problem of DT service as a Markov Decision Process (MDP).

IV. ENERGY EFFICIENT POWER CONTROL POLICY: DISCRETE CHANNEL CASE

In this section, we propose an energy efficient power control policy for the DT service with discrete channel model based on the SDP.

A. POLICY AND STATE SETTING

In general, a MDP can be defined by a 4-tuple $\langle S, A, R, Pr_t \rangle$. $S = \{s_1, s_2, ...\}$ is the states set. $A = \{a_1, a_2, ...\}$ is the action set. Meanwhile, R and Pr_t are the action reward and the state transition probability, respectively.

For the transmission power optimal problem of the DT service, the state set S is defined by the channel condition and the required information that the source node want to transmit. And, the action set A is defined by the information that the source node will transmit in current slot. Unfortunately, the required information I and channel condition γ usually are continuous variables which will lead an unsolvable bellman equation due to the curse of dimensionality. For this reason, we discrete the required information I as $(0, I/N_1, ..., (N_1 - 1)I/N_1, I)$ by a predefined integer N_1 and assume the channel condition is simply enough to be formulated as a discretization model $(\gamma_2, ..., \gamma_{N_2})$. So, the state set S is denoted as

$$s_i = (k_I * I/N_1, \gamma_{k_C}), \tag{8}$$

where $k_I \in \{0, .., N_1\}, k_c \in \{1, .., N_2\}$, and $i \in \{1, .., (N_1 + 1) \times (N_2)\}$. And, the action set A is

$$a_j = (k_a * I/N_1), \quad j \in \{1, .., N_1 + 1\}. \tag{9}$$

where $k_a \in \{0, .., N_1\}$. From the definition of S and A, the transition probability Pr_t is

$$\begin{aligned} & Pr_t(s'|s = \left\{\frac{k_I * I}{N}, \gamma_{k_c}\right\}, a = \left\{\frac{k_a * I}{N}\right\}) \\ = & \begin{cases} Pr_c(\gamma_{k_c'}), & s' = \left\{\frac{(k_s - k_a) * I}{N}, \gamma_{k_c'}\right\}; \\ 0, & others. \end{cases} \end{aligned} \tag{10}$$

where $Pr_c(\gamma_{k_c'})$ is the probability when the channel condition is $\gamma_{k_c'}$ which follows

$$\sum_{k_c'=1,2,...,N_2} Pr_c(\gamma_{k_2'}) = 1 \tag{11}$$

Furthermore, the reward $R = r_1, r_2, ..., r_T$ is the function of state and action. And, it also should reflect the transmission power $P_{t,i}$. By this way, we formulate the transmission power optimal problem of DT service as a MDP problem with finite states. Furthermore, a T-stages query table with $T \times (N_1 + 1) \times N_2 \times (N_1 + 1)$ elements can be established.

B. OPTIMAL ALGORITHM OF THE VALUE FUNCTION

Generally, the MDP problem can be solved by the Bellman equation

$$\begin{aligned} & V^t(s_i) = R(s_i) + \sum_{s_{i+1} \subseteq S} Pr_t(s_{i+1}|s_i) V^{t+1}(s_{i+1}). \\ & Q^t(s_i, a_j) = R(s_i, a_j) + \sum_{s_{i+1} \subseteq S} Pr_t(s_{i+1}|s_i, a_j) V^{t+1}(s_{i+1}). \end{aligned} \tag{12}$$

where $V^{t+1}(s_i)$ and $Q^{t+1}(s_i, a_j)$ are the state value function and state-action value function of $t + 1$ th stage, respectively. If the MDP has a terminate state, such as the transmission power optimal problem of DT service, the Bellman equation of the terminate state can be further simplified as

Policy 1: The energy efficient transmission power control policy of the DT service based on the SDP

1) Set the states and action sets
   a) Define a proper integer N.
   b) Define the state set S as (8).
   c) Define the action set A as (9).
   d) Define the state transition probability Pr_t as (10).
2) Identify the T-stages query table.
   a) Identify the state value function of the last stage by (13)(17).
   b) Back track the state value function of the other stages by (12).
   c) Identify the action of each state by (15).
3) Transmit the required information.
   a) Set the first stage as $s_1 = (I, \gamma^1_{k_c})$.
   b) loop
      i) Identify the action by the query table.
      ii) Identify the next stage based on the action.

$$V^T(s_i) = R(s_i).$$
$$Q^T(s_i, a_j) = R(s_i, a_j). \tag{13}$$

In practical, the state value function $V^t(s_i)$ can be calculated by the state-action value function $Q^t(s_i, a_j)$ as

$$V^t(s_i) = argmax\{max(Q^t(s_i, a_j)|a_j \in A)\}. \tag{14}$$

Actually, the state-action value function $Q^t(s_i, a_j)$ also indicates the best action. When the source node has the exact state-action value function, the best action of the t th stage can be identity as

$$a^t = argmax\{Q^t(s_i, a_j|a_j \in A)\}. \tag{15}$$

Based on we analyzed before, it is clear that the accurate state-action value functions are very important for the Bellman equation. Due to the constraint of the required information of DT service, the action of the last stage is

$$a^T = (k_I^T * I/N_1) \tag{16}$$

So, the last stage Bellman equation of DT service (13) should be represented as

$$\begin{aligned} Q^T(s_i, a_j) &= R(s_i, a^T) \\ &= \frac{2^{\frac{k_I^T * I}{N_1}} - 1}{\gamma^T_{k_c}} \end{aligned} \tag{17}$$

where $s_i = (k_I^T * I/N_1, \gamma^T_{k_c})$. Based on (17)(10), all of the state-action value function $Q^t(s_i, a_j)$ can be calculated as (12) by back track algorithm. Furthermore, the best action of each state can be identified as (15). The policy based on the SDP is summarized in the Policy. 1.

V. ENERGY EFFICIENT POWER CONTROL POLICY: CONTINUES CHANNEL CASE

When the channel condition is simple enough to be formulated as a discrete model, the previous algorithm is an effective method to improve the energy efficiency of the DT service. However, since the channel of the wireless communication is various and random, the discrete model is not powerful enough to depict it in most of the general case. For this reason, the Additive White Gaussian Noise (AWGN) and Multiplicative Rayleigh Fading (MRF), which are the most common channel modes of wireless communication, are considered in this section and the corresponding simulation section. Based on these assumptions, we have the following property.

Property 1. *Let the channel gain h follows the Rayleigh distribution with probability density function*

$$f(h) = \frac{h}{\sigma^2} e^{-\frac{h^2}{\sigma^2}}. \tag{18}$$

Then, the $\frac{|h|^2}{\sigma_i^2}$ follows the exponential distribution with the parameter $\frac{\sigma_i^2}{2\sigma^2}$.

Proof. See the Appendix A. □

Property. 1 shows that the channel condition γ follows the negative exponential distribution when the AWGN and MRF are adopted. So, the proposed SDP policy is not effective anymore due to the dimension curse.

A. MODEL BASED APPROACH: APPROXIMATE STATISTICAL DYNAMIC PROGRAMMING

Approximate Statistical Dynamic Programming (ASDP) is the extension of SDP. When the state value function and state-action value function cannot be calculated exactly, the ASDP is adopted to approximate the state value function and state-action value function by reducing the accuracy demand. Based on this idea, we define the state set S and action set A as $(0, I/N_1, ..., (N_1+1)I/N_1, I)$. Then, the T-stages query table has $T \times (N_1+1) \times (N_1+1)$ elements. And, the state transition probability Pr_t is

$$Pr_t(s'|s = \frac{k_s * I}{N}, a = \frac{k_a * I}{N}) = \begin{cases} 1, & s' = \frac{(k_s - k_a) * I}{N}; \\ 0, & s' \neq \frac{(k_s - k_a) * I}{N}. \end{cases} \tag{19}$$

Moreover, since the state set did not consider the CSI anymore, the Bellman equation of the terminate state (13) should be rewritten as

$$V(s) = E_\gamma[R(s)].$$
$$Q(s, a) = E_\gamma[R(s, a)]. \tag{20}$$

where $E_\gamma[\cdot]$ is the expectation about channel condition γ. Similarly, the Bellman equation of the other stage (12) is

Policy 2: The energy efficient transmission power control policy of the DT service based on the ASDP

1) Set the states and action sets
   a) Define a proper integer N.
   b) Define the state set S and the action set A as $(0, I/N_1, ..., (N_1+1)I/N_1, I)$.
   c) Define the state transition probability Pr_t as (19).
2) Identify the T-stages query table.
   a) Identify the state value function of the last stage by (20).
   b) Back track the state value function of the other stages by (21).
3) Transmit the required information.
   a) Set the first stage as $s_1 = (I, \gamma_{k_c}^1)$.
   b) loop
      i) Identify the action by the query table and (23).
      ii) Identify the next stage based on the action.

$$V(s) = E_\gamma[R(s)] + \sum_{s' \subseteq S} Pr_t(s'|s)V(s').$$
$$Q(s,a) = E_\gamma[R(s,a)] + \sum_{s' \subseteq S} Pr_t(s'|s,a)V(s'). \quad (21)$$

Different from the SDP approach which could identify the best action of each state directly, ASDP will identify the action based on the CSI of current slot. So, (15) should be modified as

$$a^t = argmax\{R(s_i, a_j, \gamma_{k_c}^t) + M(s_i, a_j) | a_j \in A)\}. \quad (22)$$

where

$$R(s_i, a_j, \gamma_{k_c}^t) = \frac{2^{s_i} - 1}{\gamma_{k_c}^t}$$
$$M(s_i, a_j) = \sum_{s_{i+1} \subseteq S} Pr_t(s_{i+1}|s_i, a_j)V^{t+1}(s_{i+1}) \quad (23)$$

The policy based on the ASDP is summarized in the Policy. 2.

B. MODEL FREE APPROACH: Q-LEARNING

When the statistical model of the continuous wireless channel is available, the ASDP is an effective method to overcome the dimension curse and improve the energy efficiency of the DT service. However, it is a challenging job to estimate the exact statistical model of the wireless channel at times. For this reason, it is more valuable to develop the model free policy for energy efficient DT service. So, the Q-learning algorithm is investigated in this section.

Based on the idea of Q-learning algorithm, the value function $V(s)$ and $Q(s,a)$ can be estimated by the transmission experience. We assume there are a set of transmission power samples

$$G(s_i, a_j) = \sum_{k=i}^{T} -P_{t,k}, \quad (24)$$

which is obtained by the transmission power control policy π in the state s_i. Then, the state-action value can be estimated as

$$\widehat{Q}_\pi^m(s_i, a_j) = \frac{1}{m} \sum_{n=1}^{m} G_n. \quad (25)$$

where $m = 1, 2, ...$, (25) can be rewritten as the recursive form

$$\widehat{Q}_\pi^m(s_i, a_j) = \widehat{Q}_\pi^{m-1}(s_i, a_j) + \frac{1}{m}(\Delta G_m(s_i, a_j)). \quad (26)$$

where

$$\Delta G_m(s_i, a_j) = G_m(s_i, a_j) - \widehat{Q}_\pi^{m-1}(s_i, a_j). \quad (27)$$

It is easy to know that the $\widehat{Q}_\pi^m(s_i, a_j)$ will converge to a exact $Q_\pi(s_i, a_j)$ with sufficient sample supportting. For this reason, a random policy is the best choice for the Q-learning algorithm to explore all of the possible states and action. On the other hand, a determine policy is preferred to pursue the best energy efficiency of DT service. In order to deal with this dilemma, the ϵ-greedy policy, as a tradeoff, is adopted. The ϵ-greedy policy can be represented as

$$\pi(a_j|s_i) = \begin{cases} \frac{\epsilon}{|A(s_i)|}, & none-greedy; \\ 1 - \epsilon + \frac{\epsilon}{|A(s_i)|}, & greedy. \end{cases} \quad (28)$$

where $|A(s_i)|$ represents the elements number of the action set when the source node is in the state s_i. Furthermore, the ϵ-greedy policy has the following property.

Property 2. *The ϵ-greedy policy with $0 < \epsilon < 1$ is an improvement and asymptotic optimal policy.*

Proof. See the Appendix B. □

(26) shows that the ϵ-greedy policy will pursue the energy efficiency with probability $1 - \epsilon + \frac{\epsilon}{|A(s_i)|}$ and explore the value function with probability $\frac{\epsilon}{|A(s_i)|}$. Furthermore, Property. 2 promise that the ϵ-greedy policy will provide the best energy efficient policy for the DT service when the $\widehat{Q}_\pi^m(s_i, a_j)$ approach to the exact $Q_\pi(s_i, a_j)$. To improve the convergency performance, the decay factor β is adopted as

$$\epsilon = \beta \times \epsilon. \quad (29)$$

The policy based on the Q-learning is summarized in the Policy. 3.

Policy. 3: The energy efficient transmission power control policy of the DT service based on the Q-learning

1) Identify a proper integer N and the 4-tuple $\langle S, A, R, Pro \rangle$.
   a) Set $S, A = (0, C/N, ..., (N-1)C/N, C)$.
   b) Initial $1 > \epsilon > 0$ and decay coefficient $1 > \beta > 0$.
   c) Initial the ϵ-greedy policy.
2) Loop forever.
   a) Generate an episode following based on the ϵ-greedy policy π: $S_0, A_0, R_1, ..., S_{T-1}, A_{T-1}, R_T$
   b) Loop for every episode.
      i) update the estimated state-action value based on (27).
      ii) update the ϵ-greedy policy as (29).

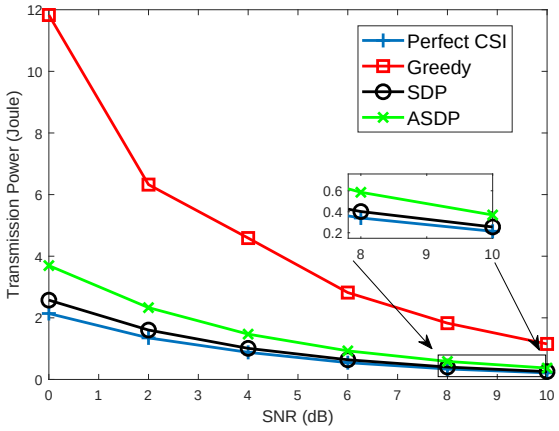


FIGURE 3. The relationship between the transmission power and the SNR

VI. SIMULATION RESULTS

In this section, the simulation results of the proposed polices are presented and analyzed. All of the results are evaluated on PC Intel (R) Core (TM) i5 CPU @ 3.20 GHz. The duration of the time slot is 1 (s). And, the required information of the destination node I is 2 (bits). We also assume the available maximal transmission power of the source node $P_{t,max}$ is 20 (Watt).

A. DISCRETE CHANNEL

In this subsection, the transmission power performance of DT service is presented when the channel fading coefficient is 0.4 or 0.9 with probability 0.5. And, the signal-noise-radio (SNR) is defined as

$$SNR = \frac{1}{\sigma^2} \tag{30}$$

In the Fig. 3, the relationship between SNR and transmission power is shown. The blue line with plus mark is the transmission power performance when the source node has the perfect future CSI. It is also the result of (7). Just as we analyzed before, this result can be regarded as the lower bound of the transmission power for the DT service. And, the red line with square denotes the transmission power performance of greedy policy. When the greedy policy

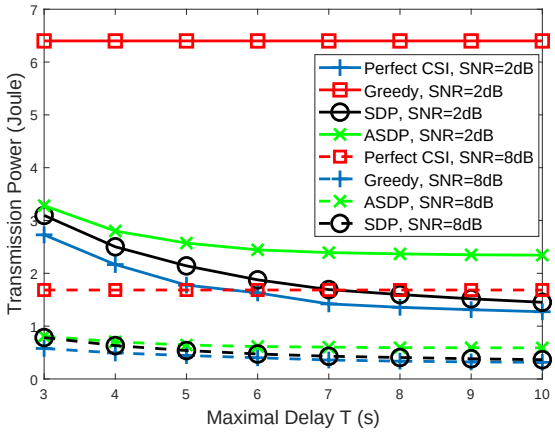


FIGURE 4. The relationship between the transmission power and the maximal tolerable delay under various SNR conditions

was adopted, the source node will work with the available maximal transmission power $P_{t,max}$ to pursue the maximal spectrum efficiency. The black line with circle and the green line with '$\times$' are the performance of SDP and ASDP policies, respectively. From the results in the Fig. 3, it is clear to see that the transmission power of the DT service decreases with the increasing of the SNR. And, the SDP policy showed the best performance which approaches to the performance upper bound effectively. Meanwhile, the performance of ASDP is slightly worse than the performance of SDP. Both of them are better than the performance of the greedy algorithm obviously. These conclusions are consistent with previous analysis. The SDP policy is developed based on the original Bellman equation without any approximation. And, the ASDP works with the reward expectation instead of the exact reward to reduce the computational complexity at the cost of slight performance degradation. Fig. 3 also showed that the performance gap among various policies decreases with the increasing of SNR. When the SNR is over 8 (dB), the performances of SDP and ASDP are very close to the performance upper bound. This result indicates that the ASDP policy is a better choice when the channel condition is good enough. On the other hand, the SDP policy is suitable for the low SNR case.

Fig. 4 presented the relationship between transmission power and the maximal tolerable delay under various SNR conditions. The solid lines are the performance when the SNR is 2 (dB). The dash lines are the performance when the SNR is 8 (dB). The meaning of the color and mark is the same as the Fig. 3. The results of Fig. 4 showed that the transmission power decreases with the increasing of the maximal tolerable delay except the greedy policy since it focuses on the current slot without considering the potential of future slots. And, the gap between the performance of greedy policy and the performance upper bound increases with the increasing of the maximal tolerable delay. These results mean that the tolerable delay of the communication service does

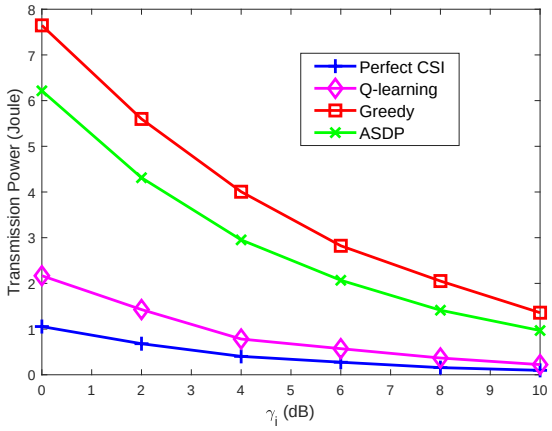


FIGURE 5. The relationship between the transmission power and γ_i

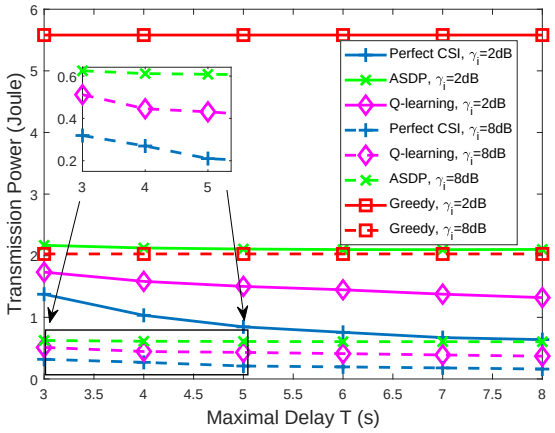


FIGURE 6. The relationship between the transmission power and the maximal tolerable delay under various γ_i conditions

provide an opportunity to decrease the transmission power. And, larger tolerable delay means a more potential for saving power.

Meanwhile, we also notice that the decreasing rate of the power lower bound also reduce with the increasing of the maximal tolerable delay. For the $SNR = 2$ (dB) case, the power lower bound is almost convergent when the maximal tolerable delay is over 7 (s). And, for the $SNR = 8$ (dB) case, the power lower bound is almost convergent when the maximal tolerable delay is over 5 (s). These results show that too long delay provides limited performance advantage for the communication service. When the channel condition is good enough, the service could finish quickly without extra power consumption.

The results of Fig. 4 also showed that the performance of SDP policy is almost the same as the power lower bound when the tolerable delay is large enough. And, the performance gap between SDP and ASDP policies increases with the increasing of the maximal tolerable delay. These results mean that the ASDP is more suitable for the DT service with smaller tolerable delay and SDP is more suitable for the DT service with larger tolerable delay.

B. CONTINUOUS CHANNEL

In this section, we assume all of the channel condition γ_i $i = 1, 2, ...T$ follow the same distribution. And, the well known Rayleigh fading channel with parameter $\sigma = \sqrt{2}$ is considered. So, from the Property. 1, the channel condition γ_i follows the exponential distribution with parameter $\frac{1}{\sigma_i^2}$.

Fig. 5 presents the relationship between γ_i and transmission power. In this case, The SDP policy is not available due to the dimension curse caused by the continuous distribution. On the other hand, the Q-learning policy, which is denoted by the pink line with diamond mark, is adopted to minimize the transmission power of the DT service. The meaning of symbols are the same as their meaning in Fig. 3. The results of Fig. 5 are similar to the results of Fig. 3. With the increasing of the γ_i, both the transmission power and the performance gaps among various policies decrease greatly. These results indicate that the service delay also provides the potential to reduce the transmission power under the continuous channel assumption. Especially in the low γ_i region, the energy efficiency potential of the DT service is very obvious. Furthermore, the performance of Q-learning policy is significantly better than the performance of ASDP policy, though it almost has no prior information about the channel. This result means that the Q-learning policy with ϵ-greedy algorithm could '*learn*' the continuous channel pretty well. We also notice that, unlike the discrete channel case, there is an obvious gap between the performance of ASDP policy and the performance upper bound though the γ_i is larger than 10 (dB). This result means the approximate of the reward leads to more performance degradation in the continuous channel.

The relationship between transmission power and maximal tolerable delay is proposed in the Fig. 6. The results show that the increasing of the maximal tolerable delay leads to the decreasing of the transmission power. And, the decline rate also decreases with the increasing of the available maximal delay. Especially in the $\gamma_i = 8$ (dB) case, the advantage of delay is very limited. We also notice that the performance of Q-learning is better than the performance of ASDP. And, there is an obvious gap between the performance of ASDP and the performance upper bound. These conclusions are similar as those of Fig. 4.

In the Fig. 7, the convergence performance of the ϵ-greedy policy is presented when the decay factor β is 0.999. For the $SNR = 2$ (dB) case, the red line with circle and blue line with plus are adopted to denote the convergence performance of ϵ-greedy policy when the maximal tolerable delays are 3 and 8 (s) respectively. Meanwhile, For the $SNR = 8$ (dB) case, the black line with '$\times$' mark and green line with square mark are adopted. From Fig. 7, it is clear to see that the transmission power of the ϵ-greedy policy decreases steadily. This result is consistent with the Property. 2 that the ϵ-greedy

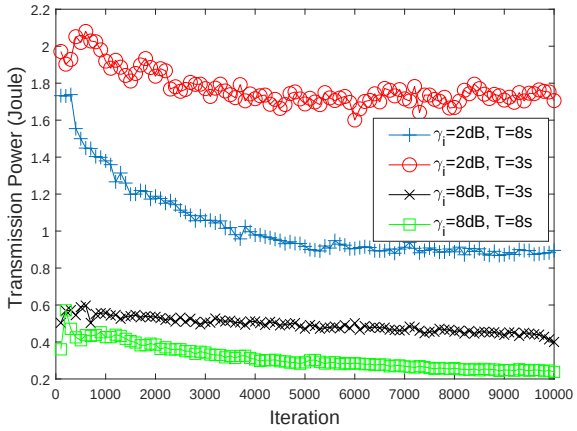


FIGURE 7. The convergence performance of the ϵ-greedy policy under various γ_i and maximal delay conditions

policy is an asymptotic optimal policy. And, Fig. 7 also show that the channel condition γ_i and maximal delay affect the convergence performance significantly. The large maximal delay leads to the large state-action space which will results in the difficulty for the ϵ-greedy policy to converge. And, the higher γ_i means the channel condition has low variance which is beneficial for the convergence performance. So, when $T = 8$ (s) and $\gamma_i = 2$ (dB), the ϵ-greedy policy needs about 5000 iterations to converge. when $T = 3$ (s) and $\gamma_i = 8$ (dB), the ϵ-greedy policy almost converged with 1000 iterations.

VII. CONCLUSION

The energy efficient communication technology plays an important role in the future wireless communication. To exploit the potential of DT service in the energy efficiency, the fundamental tradeoff between energy efficiency and delay of the DT service is investigated and analyzed. Furthermore, since it is hard to obtain perfect CSI in practical scenarios, Three energy efficient policies of the DT service is designed under various partial CSI assumptions. The numerical results show that the proposed policies could improve the energy efficiency of DT service greatly. Even if the source node has no prior information about the wireless channel, the energy consumption of DT services could still be saved by adopting the Q-learning policy. Thus, the results of this paper may provide some new insights to design energy efficient communication for DT service in the near future.

.

APPENDIX A PROPERTY 1

Let the channel gain h follows the Rayleigh distribution with PDF as

$$f(h) = \frac{h}{\sigma^2} e^{-\frac{h^2}{2\sigma^2}}. \tag{31}$$

We set a new variable as

$$y = \frac{h^2}{\sigma_i^2} \tag{32}$$

Then, we have

$$\begin{aligned} Pr(y < Y) &= Pr(\frac{h^2}{\sigma_i^2} < Y) \\ &= Pr(-\sqrt{Y\sigma_i^2} < h < \sqrt{Y\sigma_i^2}) \\ &= \int_0^{\sqrt{Y\sigma_i^2}} \frac{h}{\sigma^2} e^{-\frac{h^2}{2\sigma^2}} dh \end{aligned} \tag{33}$$

So, the PDF of y is

$$\begin{aligned} f(y) &= f(\frac{h^2}{\sigma_i^2}) \\ &= \frac{\sigma_i^2}{2\sigma^2} e^{-\frac{Y\sigma_i^2}{2\sigma^2}} \end{aligned} \tag{34}$$

(34) shows that the $\frac{h^2}{\sigma_i^2}$ follows the expential distribution with parameter $\frac{\sigma_i^2}{2\sigma^2}$.

APPENDIX B PROPERTY 2

The ϵ-greedy policy π can be written as

$$\pi(a|s) = \begin{cases} \frac{\epsilon}{A(s)}, & none-greedy; \\ 1 - \epsilon + \frac{\epsilon}{A(s)}, & greedy. \end{cases} \tag{35}$$

We define the previous ϵ-greedy policy π' as

$$\pi'(a|s) = \begin{cases} \frac{\epsilon}{A(s)} + \xi_i, & none-greedy; \\ 1 - \epsilon + \frac{\epsilon}{A(s)} - \sum_{a_i \neq a*} \xi_i, & greedy. \end{cases} \tag{36}$$

Then, we have

$$\begin{aligned} V_\pi(s) &= \sum_a \pi(a|s) Q_{\pi'}(s,a) \\ &= \frac{\epsilon}{A|s|} \sum_a Q_{\pi'}(s,a) + (1-\epsilon) \max_a Q_{\pi'}(s,a) \end{aligned} \tag{37}$$

if we denote

$$M = \max_a q_{\pi'}(s,a) \tag{38}$$

we have

$$\begin{aligned} M &\geq M - \frac{\sum_{a_i \neq a*} \xi_i}{1-\epsilon}(M - q_{\pi'}(s,a)) \\ &= \sum_a \frac{\pi(a|s) - \frac{\epsilon}{|A(s)|}}{1-\epsilon} q_{\pi'}(s,a) \\ &= M' \end{aligned} \tag{39}$$

Based on the (37)-(39), we have

$$\begin{aligned} V_\pi(s) &= \frac{\epsilon}{|A(s)|}\sum_a q_{\pi'}(s,a) + (1-\epsilon)M \\ &\geq \frac{\epsilon}{|A(s)|}\sum_a q_{\pi'}(s,a) + (1-\epsilon)M' \\ &= V_{\pi'}(s) \end{aligned} \tag{40}$$

(40) shows that the ϵ-greedy policy is improvement. Furthermore, (37) can be represented as

$$\begin{aligned} V_\pi(s) &= \sum_a \pi(a|s) Q_{\pi'}(s,a) \\ &= \frac{\epsilon}{A|s|}\sum_a Q_{\pi'}(s,a) + (1-\epsilon)\max_a Q_{\pi'}(s,a) \\ &= \frac{\epsilon}{A|s|}\sum_a \sum_{s',r} Pr_t(s',r|s,a)[r + \alpha V_\pi(s')] \\ &+ (1-\epsilon)\max_a \sum_{s',r} Pr_t(s',r|s,a)[r + \alpha V_\pi(s')] \end{aligned} \tag{41}$$

(41) also is the unique solution of the best policy. there are more detail in the [28].

REFERENCES

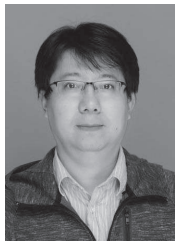
[1] SMART 2020: Enabling the Low Carbon Economy in the Information Age, The Climate Group, London, U.K., 2008.
[2] Green Power for Mobile. The Global Telecom Tower ESCO Mar- ket. [Online]. Available: http://www.gsma.com/mobilefordevelopment/wpcontent/uploads/2015/01/140617-GSMA-report-draft-vF-KR-v7.pdf.
[3] A. Fehske, G. Fettweis, J. Malmodin, and G. Biczok, "The global footprint of mobile communications: The ecological and economic perspective," IEEE Commun. Mag.., vol. 49, no. 8, pp. 55-62, Aug. 2011.
[4] Chen Y, Zhang S and Xu S et al. "Fundamental trade-offs on green wireless networks," IEEE Commun. Mag., vol. 49, no. 6, pp. 30-37, Aug. 2011.
[5] A. Abrol and R. Jha. "Power Optimization in 5G Networks: A Step Towards Green Communication," IEEE Commun. Mag., vol. 4, no. 6, pp. 1355-1374, 2016.
[6] R. Mahapatra, Y. Nijsure and G. Kaddoum et al., "Energy Efficiency Tradeoff Mechanism Towards Wireless Green Communication: A Survey," IEEE Communications Surveys and Tutorials., vol. 18, no. 1, pp. 686-705, Firstquarter 2016.
[7] P. Gandotra, R. K. Jha and S. Jain, "Green Communication in Next Generation Cellular Networks: A Survey," IEEE Access, vol. 5, pp. 11727-11758, 2017.
[8] Alcatel-Lucent and Vodafone Chair on Mobile Communication Systems: Study on energy efficient radio access network (EERAN) technologies, Tech. Univ. Dresden, Dresden, Germany, Project Rep., 2009.
[9] Y. Soh, T. Quek, M. Kountouris, and H. Shin, "Energy efficient heterogeneous cellular networks," IEEE J. Sel. Areas Commun., vol. 31, no. 5, pp. 840-850, May 2013.
[10] S. Zhou et al., "Green mobile access network with dynamic base station energy saving," ACM MobiCom, vol. 9, no. 262, pp. 10-12, 2009.
[11] J. Gong, S. Zhou, Z. Niu, and P. Yang, "Traffic-aware base station sleeping in dense cellular networks," in Proc. IEEE 18th Int. Workshop Quality Service, Boston, USA, 2010, pp. 1-2.
[12] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz, "Primary user behavior in cellular networks and implications for dynamic spectrum access," IEEE Commun. Mag., vol. 47, no. 3, pp. 88-95, Mar. 2009.
[13] L. Suarez, L. Nuaymi, and J.-M. Bonnin, "Analysis of a green-cell breathing technique in a hybrid access network environment," in Proc. IFIP Wireless Days, Valencia, Spain, 2013, pp. 1-6.
[14] R. Torrea-Duran, P. Tsiaakis, L. Vandendorpe, and M. Moonen, "A cell breathing approach in green heterogeneous networks," in Proc. IEEE 15th Int. Workshop Signal Process. Adv. Wireless Commun., Toronto, Canada, 2014, pp. 344-348.
[15] H. Wang et al., "Liquid cell management for reducing energy consumption expenses in hybrid energy powered cellular networks," in Proc. IEEE Wireless Commun. Netw. Conf., Marrakesh, Morocco, 2014, pp. 1632-1637.
[16] Y. Xu, H. Li and Z. Feng et al., "Energy sustainability modeling and liquid cell management in green cellular networks," in Proc. IEEE Int. Conf. Commun., budapest, hungary, 2013, pp. 4414-4419.
[17] H. Goudarzi and M. Pakravan, "Equal power allocation scheme for cooperative diversity," in Proc. 4th IEEE/IFIP Int. Conf., Central Asia, 2008, pp. 1-5.
[18] M. Lau and W. Yue, "Optimality and feasibility of equal power allocation of IDMA systems," in Proc. 5th Int. Symp. Modeling Optim. Mobile, Ad Hoc Wireless Netw. Workshops, Limassol, Cyprus, 2007, pp. 1-7.
[19] J. Jang, K. Lee, and Y. Lee, "Transmit power and bit allocations for OFDM systems in a fading channel," in Proc. IEEE Global Telecommun. Conf., Waikoloa, USA, 2003, pp. 1-7.
[20] X. Gong, A. Ispas, G. Dartmann, and G. Ascheid, "Power allocation and performance analysis in spectrum sharing systems with statistical CSI," IEEE Trans.Wireless Commun., vol. 12, no. 4, pp. 1819-1831, Apr. 2013.
[21] M. Mili, L. Musavian, K. Hamdi and F. Marvasti, "How to increase energy efficiency in cognitive radio networks," IEEE Trans. Commun., vol. 64, no. 5, pp. 1829-1843, May 2016.
[22] Q. Wu, W. Chen and M. Tao et al., "Resource allocation for joint transmitter and receiver energy efficiency maximization in downlink OFDMA systems," IEEE Trans. Commun., vol. 63, no. 2, pp. 416-430, Feb. 2015.
[23] L. Zhang, M. Xiao and G. Wu et al., "Efficient scheduling and power allocation for D2D-assisted wireless caching networks," IEEE Trans. Commun., vol. 64, no. 6, pp. 2438-2452, Jun. 2016.
[24] Y. Luo, P. Hong, and R. Su, "Energy-efficient scheduling and power allocation for energy harvesting-based D2D communication," in Proc. IEEE Glob. Commun. Conf., Singapore, 2017, pp. 1-6.
[25] I. Ahmed, K. Phan, and T. Le-Ngoc, "Optimal stochastic power control for energy harvesting systems with delay constraints," IEEE J. Sel. Areas Commun., vol. 34, no. 12, pp. 3512-3527, Dec. 2016.
[26] Y. Cui, V. Lau, and F. Zhang, "Grid power-delay tradeoff for energy harvesting wireless communication systems with finite renewable energy storage," IEEE J. Sel. Areas Commun., vol. 33, no. 8, pp. 1651-1666, Aug. 2015.
[27] W. Huang, W. Chen, and H. Poor, "Energy efficient wireless pushing with request delay information and delivery delay constraint," IEEE Access, vol. 5, pp. 15428-15441, 2017.
[28] R. Sutton and A. Barto, Reinforcement Learning: An Introduction, Jan. 2018.
[29] W. Powell, "Approximate Dynamic Programming: Solving the Curses of Dimensionality (Wiley Series in Probability and Statistics)," optimization methods and software, vol. 24, no. 1, pp. 155-155, 2007.
[30] W. Powell, Approximate Dynamic Programming, Nov. 2011.
[31] Y. Luo, M. Zeng and H. Jiang, "Learning to Tradeoff Between Energy Efficiency and Delay in Energy Harvesting-Powered D2D Communication: A Distributed Experience-Sharing Algorithm," IEEE Internet of Things Journal, vol. 6, no. 3, pp. 5585-5594, June 2019.
[32] K. Nguyen, T. Duong and N. Vien et al., "Non-Cooperative Energy Efficient Power Allocation Game in D2D Communication: A Multi-Agent Deep Reinforcement Learning Approach," IEEE Access, vol. 7, pp. 100480-100490, 2019.

RUI ZHU received the B.S. and M.S. degrees in electronic engineering from the PLA University of Science and Technology, Nanjing, China, in 2002 and 2005,respectively. And he received the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2014.

From 2005 to 2007, he was a Research Assistant with the Telecommunication Engineering Institute of the Air Force. From 2008 to 2010, he was a lecturer with the Telecommunication Engineering Institute of the Air Force. From 2014 to 2017, he was a lecturer with the Information and Navigation Institute, Air Force Engineering University. Since 2018, he has been an assistant professor with the Information Engineering School, Xijing University. He is the author of more than 40 articles,. His research interests include green communication, cognitive radio network and wireless resource allocation.

JIANXIN GUO received the B. S. degree in Communications Engineering from Telecommunications Engineering Institute of the Air Force, Xi'an, China, in 1997, the M. S. degree in Information and Communications Engineering from Air Force Engineering University, Xi'an, China, in 2000, and the Ph.D. degree in Information and Communications Engineering from the PLA Information Engineering University, Zhengzhou, China, in 2004.

From 2005 to 2007, he was a lecturer with the Telecommunication Engineering Institute of Air Force. From 2008 to 2009, he was an assistant professor with the Telecommunication Engineering Institute of Air Force. From 2009 to 2017, he was a professor with the Information and Navigation Institute, Air Force Engineering University. Since 2018, he has been a professor with the Information Engineering School, Xijing University. He is the author of two books, more than 30 articles,. His research interests include green communication, cognitive radio network and wireless resource allocation.



FENG WANG received the B. S. degree in Communications Engineering from Air Force Engineering University (AFEU), Xiąfan, China, in 2001,and the Ph.D. degree in Information and Communications Engineering from the Xidian University Xiąfan, China, in 2016.

From 2004 to 2018, he was a lecturer with the AFEU. Since 2018, he has been an assistant professor with Information Engineering School, Xijing University. His research interests include compressed sensing, wireless localization and optimization theory.

BAOQIN LIN received the M.Sc. degree in electromagnetic field and microwave technology from Air Force Engineering University, Xiąfan, China, in 2002, and the Ph.D. degree in Electronic Science and Technology from National University of Defense Technology, Changsha, China, in 2006. He is currently an Associate Professor with the Department of Information Engineering, Xijing University, Xiąfan, China.

From 2002 to 2006, he was a Research Assistant with the Telecommunication Engineering Institute of the Air Force. From 2007 to 2016, he was a lecturer with the Telecommunication Engineering Institute of the Air Force. Since 2017, he has been an assistant professor with the Information Engineering school, Xijing University. He is the author of more than 30 articles,. His research interests include wireless communication, phased array antennas and ultra-wideband antennas. He is a member of IEEE Antennas and Propagation Society.

YANGCHAO HUANG rereceived the B. S. degree in Communications Engineering from Telecommunications Engineering Institute of the Air Force, Xi'an, China, in 2001, the M. S. degree in Information and Communications Engineering from Air Force Engineering University, Xi'an, China, in 2004, and the Ph.D. degree in Information and Communications Engineering from the PLA Information Engineering University, Xi'an, China, in 2016.

From 2004 to 2006, he was a Research Assistant with the Telecommunication Engineering Institute of the Air Force. From 2007 to 2016, he was a lecturer with the Telecommunication Engineering Institute of the Air Force. Since 2017, he has been an assistant professor with he Information and Navigation Institute, Air Force Engineering University. He is the author of more than 30 articles,. His research interests include green communication, cognitive radio network and communication counter.