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

RUI ZHU¹, GUO JIANXIN¹, WANG FENG², BAOQIN LIN³, (Member, IEEE), AND YANGCHAO HUANG²

¹School of Information Engineering, Xijing University, Xi’an 710123, China
²Institute of Information and Navigation, Air Force Engineering University, Xi’an 710008, China

Corresponding author: Wang Feng (wangfengisu@163.com)

This work was supported in part by the Nature Science Foundation of Shaanxi Province under Grant 2018JM6075, in part by the Shaanxi Key Laboratory of Integrated and Intelligent Navigation Open Fund under Grant SKLIIIN-20180211, in part by the Research Foundation for Talented Scholars of Xijing University under Grant XJJ19B01 and Grant XJJ17B06, in part by the Key Research and Development Plan Project of Shaanxi Provincial Science and Technology Department under Grant 2018ZDXXM-NY-014, and in part by the Xi’an Science and Technology Project under Grant 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 spectrum efficiency provides more opportunity to save power by adjusting the transmission power of 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) 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 commercial deployment of the fifth generation (5G) wireless communication networks around the world, all of the researcher consider what the sixth generation (6G) wireless communication network should be. Although the specific standard of 6G is in the preliminary period, it is no doubt that tens of billions smart devices, such as sensors and wearable devices, will be enabled to interact with each other to fulfill the vision of connecting everything worldwide in the future 6G era. This trend will lead to an explosive growth of energy demand and carbon emissions while offering advantage [1], [2]. Energy consumption and carbon emission will not only degrade the quality of environment acutely but also increase 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 became a popular topic.

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 is considered as an effective
way to prolong the live life of power-constraint terminal devices.

In fact, technologies towards energy-efficient wireless networks have been studied for a long time. 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. simplified 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], serial indices of the energy efficiency were proposed in terms of network convergence, throughput and spectrum efficiency. With the maturity of the research frame of the energy efficient communication, the specific technology was investigated to realize the energy efficient communication.

In this paper, we investigated the energy efficiency potential of the Delay Tolerable (DT) service with partial Channel States Information (CSI). We also proposed three transmission power control policies under various assumptions. 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 CSI, which may not be available in practice. In order to improve the utility value of research, we develop a model free transmission policy to improve the energy efficiency of DT service based on the Q-learning algorithm.

The remainder of this paper is organized as follows. In Section II, the related works of energy efficient communication are briefly 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

Energy efficient power control and allocation is a challenging problem in modern wireless network due to the complex network structure and various service constraints. Many researchers contributed substantially to this field. In [8], the sleep mechanism was proposed to control the static power by components deactivation when they are not working. Following this idea, [9]–[11] optimized the mode switching policy of sleep mechanism based on the loading condition and traffic intensity with the blocking probability constraint. Furthermore, with the explosive rise in the number of subscribers and traffic, the concept of cell breathing emerged as the generalized BS sleeping mode in [12]. The cell breathing was more flexible and effective since it can adjust the cell size in accordance with the loading and QoS variations. The conclusions of [13], [14] showed that the cell breathing provided excellent energy efficiency performance gain in the complex cellular network.

Besides the static power, the optimization of transmission power is also generally investigated. In [15], the double-threshold water filling (WF) algorithm was proposed to improve the spectrum efficiency and energy efficiency in the spectrum sharing systems with the statistical CSI. In [16], a multi-objective optimal problem was formulated to optimize the energy efficiency and spectrum efficiency of the cognitive radio networks. By the $\epsilon$-constraint method, the minimum power allocation strategy of the secondary transmitter was obtained while the average interference power on the primary user receiver was limited. In [17], a joint optimal problem of the OFDMA system was formulated. And, the researcher also proposed an iterative algorithm to solve the proposed optimal problem. In [18], a link adaptive power control and allocation (LaPCA) scheme was designed based on the requirement of LTE service. Instead of optimizing the energy efficiency directly, LaPCA kept the portion of cell transmission power to be proportional to the volume of data flows that were nominated for transmission. In order to overcome the shortage of centralized approaches, game theoretic has attracted significant attention to develop distributed power control and allocation approaches. In [19], [20], the problem of QoS-driven power control in the uplink of CDMA wireless networks is formulated as a non-convex noncooperative Multi-Service Uplink Power Control (MSUPC) game. Based on the diverse user’s
requirements, various utility functions were designed with respect to the throughput performance, QoS requirements fulfillment and the power consumption. And, the corresponding distributed iterative algorithm were also proposed and evaluated. In [21], a simulation-based optimization framework for D2D network was proposed to achieve a tradeoff between the energy consumption and network performance. Although the existence of Nash equilibrium point was not proved theoretically, the simulation results showed that the proposed optimization framework has an excellent performance in most case.

In all of the related research, the tradeoff between energy efficiency and service delay kept playing an important role. In [22], [23], the relationship between energy efficiency and service delay was analyzed and simulated. The conclusion showed that the energy efficiency and delay were contradictory to each other. In [24], the time-domain power allocation scheme was studied to improve the energy efficiency of the D2D system in the delay-insensitive scenario. In [25], the mechanism of pro-actively pushing content to users was adopted to reduce the energy consumption due to the extended transmission time duration. Furthermore, the user request probability thresholds was derived under two different QoS constraints in this paper. In [26], the grid power-delay tradeoff of the energy harvesting wireless system with finite energy storage capacity was studied. The bounds of the average delay and grid power consumption was derived. In order to improve the utility value of the relative research, an energy efficiency and delay tradeoff optimization scheme was proposed under the limited energy and data storage space assumption [27]. A modified distributed Q-learning algorithm was also developed to cooperative with the proposed optimization scheme in this paper. The simulation results showed that the proposed algorithm and scheme can obtain the near-optimal performance as the classical centralized Q-learning mechanism.

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 provided an effective tool to improve the performance of wireless communication with partial CSI [28]–[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 time-slotted communication link with 2 nodes [32] is studied in this paper. A time slot is a fixed time interval and could consist of one or several packets. Meanwhile, the multiplicative fading and the additive white Gaussian noise (AWGN) are adopted. Without loss of generality, we assume that the fading gains of the channel between the users are affected by fast fading, shadow fading, and long-time scale variations and thus, can be modeled as a stationary time-varying stochastic process. Furthermore, we also assume the channel gain within the duration of each time slot is constant. So, the receiving signal of the destination node in the $i$ the slot can be presented as

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

where $P_{t,i}$ (Watt) is the transmission power of $i$ th slot. It is fixed within the duration of a time 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. Unlike the delay sensitive service, the DT service requires a certain amounts information in a period rather than immediately. For example, the video-on-demand (VOD) is a typical DT service when the spectrum efficiency is good enough. The video will keep playing smoothly if the new data can fill the buffer before it is NULL. So, the maximal tolerable delay of VOD depends on the video playing rate and the buffer pool size of terminal node. And, the energy efficiency of DT service can be improved by optimizing the transmission power in time domain. The Fig. 2 illustrate the optimal power allocation pattern for the DT service which requires to transfer $I$ bits in 8 slots. The solid bars in the Fig. 2 represent the channel condition reciprocal of the wireless channel which is defined as

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

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.

When the energy efficient communication is studied, a popular and precise indices is the energy efficiency which

![FIGURE 1. An illustration of the point to point communication link with a source node and a destination node.](image-url)
The power constraint is given by \( \gamma \), where \( \sum_{i} \) represents the total information transmitted in a period. In the following part, we will focus on the minimization of the total 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. In general case, the higher energy efficiency means the source node can transmit more information with limited power. However, when we consider a DT service with the certain amount of information and maximal tolerable delay, such as VOD or file transfer, the energy efficiency problem (3) can be modified as

\[
\frac{I}{W(\tau P_f + \sum_{i=1}^{T} P_{t,i})} \text{ (bit/Joule/Hz),}
\]

where \( I \) (bits) and \( T \) are the required total information amount and the maximal number of tolerable delay slots respectively. \( \tau \) (s) is the duration of slot. And, \( P_{t,i} \) (Watt) is the transmission power in the \( i \)th slot as we defined before. Different from the traditional energy efficiency, the information amount \( I \) is a constant. And, since the receive node require to receive all of these information in a period, the total static power \( W \tau P_f \) is also a constant. Then, the optimal problem of energy efficiency can be turned to the optimal problem of transmission power \( W \tau \sum_{i=1}^{T} P_{t,i} \). In order to reduce the analysis complexity, \( \tau \) and \( W \) are normalized in this paper. So, in the following part, we will focus on the minimization problem of \( \sum_{i=1}^{T} P_{t,i} \).

### B. Problem Formulated

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

\[
I = \sum_{i=1}^{T} \log_2(1 + P_{t,i} \gamma_i),
\]

where \( \gamma_i \) is CSI which is defined in (2). And, the transmission power constraint is

\[
P_{t,i} \leq P_{t,i,\text{max}}, \quad i = 1, 2, \ldots, T.
\]

where \( P_{t,i,\text{max}} \) is the available maximal transmission power of source node in the \( i \)th slot. So, the transmission power optimal problem of the DT service can be formulated as

\[
\begin{align*}
\min & : \sum_{i=1}^{T} P_{t,i} \\
\text{s.t.} & : \sum_{i=1}^{T} \log_2(1 + P_{t,i} \gamma_i) \geq I, \\
& : P_{t,i} \leq P_{t,i,\text{max}}, \quad i = 1, 2, \ldots, T. \quad (7)
\end{align*}
\]

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 (\( \gamma_i \)). 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. On the other hand, with the development of various channel estimation, the source node can get a pretty well CSI of the current slot and the statistical CSI of the future slots. This advantage means the source node should adjust the transmission power base on the observed CSI. Following this idea, we reformulated the transmission power optimal problem of DT service as a Markov Decision Process (MDP).

### IV. Energy Efficient Power Control Policy for the Discrete Channel

In this section, we propose an energy efficient power control policy for the DT service with discrete channel model based on the SDP. And, as we mentioned before, the DT service requires a certain amounts information \( I \) in a period \( T \) rather than immediately. So, the state set of policy can be designed as the following subsection.

#### A. Policy State Setting

In general, a MDP can be defined by a 4-tuple \( (S, A, R, Pr_t) \). \( S = \{s_1, s_2, \ldots \} \) is the states set. \( A = \{a_1, a_2, \ldots \} \) 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 CSI 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 CSI (\( \gamma \)) usually are continuous variables which will lead to an unsolvable bellman equation due to the curse of dimensionality. For this reason, we discretize the required information \( I \) as \( (0, I/N_1, \ldots, (N_1 - 1)I/N_1, I) \) by a predefined integer \( N_1 \) and assume the CSI is simply enough to be formulated as a discretization model (\( \gamma_1, \ldots, \gamma_{N_2} \)) in this section. So, the state set \( S \) is denoted as

\[
s_i = (k_I * I/N_1, \gamma_{k_C}), \quad (8)
\]
where \( k_1 \in \{0, ..., N_1\}, k_c \in \{1, ..., N_2\}, \) and \( i \in \{1, ..., (N_1 + 1) \times (N_2)\} \). Meanwhile, the action set \( A \) is

\[
a_j = (k_a * I / N_1), \quad j \in \{1, ..., N_1 + 1\},
\]

(9)

where \( k_a \in \{0, ..., N_1\} \). From the definition of \( S \) and \( A \), the transition probability \( P_{t}\) is

\[
Pr_t \left( s' | s = \left( k_t * I / N_1, \gamma_k \right), a = \left( k_a * I / N_1 \right) \right) = \begin{cases} 
Pr_c(\gamma_k'), \quad s' = \left( (k_a - k_c) * I / N_1, \gamma_k' \right); \\
0, \quad \text{others.}
\end{cases}
\]

(10)

where \( Pr_c(\gamma_k') \) is the probability when the CSI is \( \gamma_k' \) which follows

\[
\sum_{k' = 1, 2, ..., N_2} Pr_c(\gamma_k') = 1
\]

(11)

Furthermore, the reward \( R = r_1, r_2, ..., r_T \) is the function of state and action. And, it also 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 can be established.

**B. OPTIMAL ALGORITHM OF THE VALUE FUNCTION**

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

\[
V^i(s_i) = R(s_i) + \sum_{s_{i+1} \in S} Pr_t(s_{i+1} | s_i)V^{i+1}(s_{i+1}).
\]

\[
Q^i(s_i, a_j) = R(s_i, a_j) + \sum_{s_{i+1} \in S} Pr_t(s_{i+1} | s_i, a_j)V^{i+1}(s_{i+1}).
\]

(12)

where \( V^{i+1}(s_i) \) and \( Q^{i+1}(s_i, a_j) \) 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

\[
V^T(s_i) = R(s_i).
\]

\[
Q^T(s_i, a_j) = R(s_i, a_j).
\]

(13)

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

\[
V^i(s_i) = \arg \max \{ \max (Q^i(s_i, a_j) | a_j \in A) \}.
\]

(14)

Actually, the state-action value function \( Q^i(s_i, a_j) \) 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

\[
a^t = \arg \max \{ Q^i(s_i, a_j | a_j \in A) \}.
\]

(15)

So, 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

\[
a^T = (k_T^T * I / N_1)
\]

(16)

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

\[
Q^T(s_i, a_j) = R(s_i, a^T)
= \frac{2^{\frac{\gamma_k^2}{\sigma^2}} - 1}{\gamma_k^2}
\]

(17)

where \( s_i = (k_T^T * I / N_1, \gamma_k^T) \). Based on (10), (12) and (17), all of the state-action value function \( Q^i(s_i, a_j) \) can be calculated by the back track algorithm. Furthermore, the best action of each state can be identified by (15). The policy based on the SDP is summarized in the Policy. 1.

---

**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 \( P_{t, i} \) 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 \( s_1 = (I, \gamma_k^i) \).
   b) loop
      i) Identify the action by the query table.
      ii) Identify the next stage based on the action.

---

**V. ENERGY EFFICIENT POWER CONTROL POLICY FOR THE CONTINUES CHANNEL**

When the CSI 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. Meanwhile, we have the following property.

**Property 1:** Let the channel gain \( h \) follows the Rayleigh distribution with probability density function

\[
f(h) = \frac{h}{\sigma^2} e^{\frac{-h^2}{\sigma^2}}.
\]

(18)
Then, the $\frac{|h|^2}{\sigma^2}$ follows the exponential distribution with the parameter $\frac{\sigma^2}{2\gamma}$. 

**Proof:** See the Appendix A. □

Property 1 shows that the CSI $\gamma$ 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 POLICY BY 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. Following this idea, we define the state set $S$ and action set $A$ as $(0, I/N_1, \ldots, (N_1 - 1)/N_1, I)$. Then, the T-stages query table has $T \times (N_1 + 1)$ elements. And, the state transition probability $P_{ri}$ is

$$
Pr_i \left(s' | s = \frac{k_i \ast I}{N_1}, a = \frac{k_a \ast I}{N_1} \right) = \begin{cases} 
1, & s' = \frac{(k_a - k_i) \ast I}{N_1} \\
0, & s' \neq \frac{(k_a - k_i) \ast I}{N_1}.
\end{cases}
$$

(19)

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

$$
V(s) = E_{\gamma}[R(s)].
$$

(20)

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

(21)

where $E_{\gamma}[\cdot]$ is the expectation about CSI $\gamma$. Similarly, the Bellman equation of the other stage (12) is

$$
V(s) = E_{\gamma}[R(s)] + \sum_{s' \in S} Pr_i(s' | s)V(s').
$$

$$
Q(s, a) = E_{\gamma}[R(s, a)] + \sum_{s' \in S} Pr_i(s' | s, a)V(s').
$$

(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

$$
a' = \arg \max \{R(s_i, a_j, y'_{k_e}) + M(s_i, a_j) | a_j \in A \}.
$$

(22)

where

$$
R(s_i, a_j, y'_{k_e}) = \frac{2g_i - 1}{y'_{k_e}}
$$

$$
M(s_i, a_j) = \sum_{s_{i+1} \in S} Pr_i(s_{i+1} | s_i, a_j)V^{t+1}(s_{i+1})
$$

(23)

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

**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 $(0, I/N_1, \ldots, (N_1 - 1)/N_1, I)$.
   c) Define the state transition probability $Pr_i$ 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 $s_1 = (I, y'_{k_e})$.
   b) loop
      i) Identify the action by the query table and (23).
      ii) Identify the next stage based on the action.

**B. MODEL FREE POLICY BY 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

$$
G(s_i, a_j) = \sum_{k=i}^{T} P_{r_{i,k}},
$$

(24)

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

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

(25)

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

$$
\hat{Q}^m(\pi, s_i, a_j) = \hat{Q}^{m-1}(\pi, s_i, a_j) + \frac{1}{m}(\Delta G_m(s_i, a_j)),
$$

(26)

where

$$
\Delta G_m(s_i, a_j) = G_m(s_i, a_j) - \hat{Q}^{m-1}(\pi, s_i, a_j).
$$

(27)

It is easy to know that the $\hat{Q}^m(\pi, s_i, a_j)$ will converge to a exact $Q(\pi, s_i, a_j)$ with sufficient sample supporting. 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 \( \epsilon \)-greedy policy, as a tradeoff, is adopted. The \( \epsilon \)-greedy policy can be represented as

\[
\pi(a|s) = \begin{cases} 
\epsilon & \text{none - greedy;} \\
1 - \epsilon + \frac{\epsilon}{|A(s)|}, & \text{greedy.}
\end{cases}
\] (28)

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

Property 2: The \( \epsilon \)-greedy policy with \(0 < \epsilon < 1\) is an improvement and asymptotic optimal policy.

Proof: See the Appendix B. \( \square \)

(26) shows that the \( \epsilon \)-greedy policy will pursue the energy efficiency with probability \( 1 - \epsilon + \frac{\epsilon}{|A(s)|} \) and explore the value function with probability \( \frac{\epsilon}{|A(s)|} \). Furthermore, Property 2 promise that the \( \epsilon \)-greedy policy will provide the best energy efficient policy for the DT service when the \( \hat{Q}_n^\epsilon(s, a) \) approach to the exact \( Q_n(s, a) \). To improve the convergency performance, the decay factor \( \beta \) is adopted as

\[ \epsilon = \beta \times \epsilon. \] (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 \( (S, A, R, \rho) \).
   a) Set \( S, A = (0, C/N, \ldots, (N - 1)C/N, C) \).
   b) Initial \( I > \epsilon > 0 \) and decay coefficient \( I > \beta > 0 \).
   c) Initial the \( \epsilon \)-greedy policy.
2) Loop forever.
   a) Generate an episode following based on the \( \epsilon \)-greedy policy \( \pi: S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T \)
   b) Loop for every episode.
      i) update the estimated state-action value based on (27).
      ii) update the \( \epsilon \)-greedy policy as (29).

C. IMPLEMENTATION COST AND COMPLEXITY ANALYSIS

From the above mentioned, we can see that the energy efficient power control policy can be developed based on the SDP, ASDP and Q-learning algorithm. In general, the statistical CSI, which is required by SDP and ASDP, is obtained by channel estimating technology in the realistic wireless communication environment. So, the implementation costs of SDP and ASDP mainly depend on the cost of adopted channel estimating technology. Meanwhile, for the Q-Learning, the transmission experience is required to estimate the value function. So, the implementation cost of Q-Learning mainly lies in the time cost which is used to sample the transmission experience. In the following, we will conclude the time and space complexity of the various energy efficient power control policies.

The state space of SDP is based on the predefined integer \( N_1 \), channel state amount \( N_2 \) and maximal tolerable delay \( T \). So, the states space complexity is \( O(N_1N_2T) \). The action space of SDP is same as the state space. So, the overall space complexity of the energy efficient power control policy based on the SDP is \( O(2N_1N_2T) \).

The time complexity is corresponding to the space complexity. In order to identify the query table, the source node has to evaluate the state value function and state-action value function among various states. So, the corresponding time complexity is \( O(N_1^2N_2^2T) \). Meanwhile, during the transmission stage, the policy is a typical lookup table algorithm. So, the corresponding time complexity is \( O(N_1N_2T) \).

Different from the SDP, the space complexity of ASDP only depends on the predefined integer \( N_1 \) and the maximal tolerable delay \( T \). So, the states space complexity is \( O(N_1T) \). Meanwhile, there is no action space for the ASDP. So, the overall space complexity of the energy efficient power control policy based on the ASDP is the states space complexity \( (O(N_1T)) \). Similarly, the time complexity of query table building is \( O(N_1^2T) \). Meanwhile, during the transmission stage, the source node should identify the action based on the CSI of current slot. So, the corresponding time complexity is \( O(N_1T) \).

The space complexity of Q-learning is same as the ASDP. The time complexity depends on the convergence rate. If we denote \( K \) as the convergence iteration times, the time complexity of Q-learning is \( O(KN_1T) \).

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 each slot \( P_{t,i,max} \) is same as 20 (Watt).

A. NUMERICAL RESULTS UNDER DISCRETE CHANNEL ASSUMPTION

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

$$\text{SNR} = \frac{1}{\sigma^2}$$

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 was adopted, the source node will work with the available maximal transmission power $P_{t,i,max}$ to pursue the maximal spectrum efficiency. The black line with circle and the green line with ‘×’ 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 ASFP 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 ASFP 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 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 $\text{SNR} = 2$ (dB) case, the power lower bound is almost convergent when the maximal tolerable delay is over 7 (s). And, for the $\text{SNR} = 8$ (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 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. Numerical Results under Continuous Channel Assumption

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

Fig. 5 presents the relationship between $\gamma_i$ 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 $\gamma_i$, 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 $\gamma_i$ 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 $\epsilon$-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 $\gamma_i$ 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 $\epsilon$-greedy policy is presented when the decay factor $\beta$ is 0.999. For the $\text{SNR} = 2$ (dB) case, the red line with circle and blue line with plus are adopted to denote the convergence performance of $\epsilon$-greedy policy when the maximal tolerable delays are 3 and 8 (s) respectively. Meanwhile, For the $\text{SNR} = 8$ (dB) case, the black line with ‘×’ mark and green line with square mark are adopted. From Fig. 7, it is clear to see that the transmission power of the $\epsilon$-greedy policy decreases steadily. This result is consistent with the Property. 2 that the $\epsilon$-greedy policy is an asymptotic optimal policy. And, Fig. 7 also show that the channel condition ($\gamma_i$) and maximal delay affect the convergence performance significantly. The large maximal delay leads to the large state-action space which will results in the convergence difficulty for the $\epsilon$-greedy policy. And, the higher $\gamma_i$ 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 $\epsilon$-greedy policy needs about 5000 iterations to converge. when $T = 3$ (s) and $\gamma_i = 8$ (dB), the $\epsilon$-greedy policy almost converged with 1000 iterations.

C. RUNNING TIME

In this section, the running time of the proposed power control policies is evaluated when the maximal tolerable
TABLE 2. Running time of various energy efficient power control policies.

|          | SDP(s) | ASDP(s) | Q-Learning(s) |
|----------|--------|---------|---------------|
|          | Prepare Stage | Transmission Stage | Prepare Stage | Transmission Stage | Transmission Stage |
| Discrete Channel | 3.49 \times 10^{-4} | 4.12 \times 10^{-6} | 1.67 \times 10^{-4} | 1.16 \times 10^{-5} | / |
| Containers Channel | / | / | 4.17 \times 10^{-2} | 1.56 \times 10^{-5} | 1.54 \times 10^{-1} |

delay (T) is 8 (s). Just as we mentioned above, the results are evaluated on PC Intel (R) Core (TM) i5 CPU @ 3.20 GHz with MATLAB 2017. And, the SNR of discrete channel and the γ of continues channel are 8(dB). The running time of ASDP and SDP is divided into the prepare stage time and transmission stage time. The prepare stage time denotes the running time which is used to identity the query table. And, the transmission stage time spends in the communication process. From the results in the Table 2, we can find that the prepare stage time of SDP is larger than it of the ASDP. Meanwhile, the transmission stage time of SDP is less than it of ASDP since the ASDP has to identify the action based on the channel condition of current slot. We also notice that the time efficiency of the proposed policies is excellent. Even when the iteration number of Q-learning K is 10000, the running time is less than 0.2(s). The results suggest that the proposed policies are applicable for real-time application.

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 when 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

\[
Pr(y < Y) = Pr\left(\frac{h^2}{\sigma_i^2} < Y\right) = Pr\left(-\frac{\sqrt{Y\sigma_i^2}}{\sigma_i} < h < \sqrt{Y\sigma_i^2}\right) = \int_0^{\sqrt{Y\sigma_i^2}} \frac{h}{\sigma^2} e^{-\frac{h^2}{2\sigma^2}} dh \tag{33}\]

So, the PDF of y is

\[
f(y) = f\left(\frac{h^2}{\sigma_i^2}\right) = \frac{1}{2\sigma_i^2} e^{-\frac{y\sigma_i^2}{2\sigma_i^2}} \tag{34}\]

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

APPENDIX B

PROPERTY 2
The \( \epsilon \)-greedy policy \( \pi \) can be written as

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

We define the previous \( \epsilon \)-greedy policy \( \pi' \) as

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

Then, we have

\[
V_\pi(s) = \sum_a \pi(a|s)Q_\pi(s, a) = \frac{\epsilon}{A(s)} \sum_a Q_\pi(s, a) + (1 - \epsilon) \max_a Q_\pi(s, a) \tag{37}\]

if we denote

\[
M = \max_a q_\pi(s, a) \tag{38}\]

we have

\[
M \geq M' + \frac{-\sum_{a_i \neq a} \xi_i (M - q_\pi(s, a))}{1 - \epsilon} = \sum_a \pi(a|s) - \frac{\epsilon}{|A(s)|} q_\pi(s, a) = M' \tag{39}\]

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

\[
V_\pi(s) = \frac{\epsilon}{|A(s)|} \sum_a q_\pi(s, a) + (1 - \epsilon)M \geq \frac{\epsilon}{|A(s)|} \sum_a q_\pi(s, a) + (1 - \epsilon)M' = V_\pi'(s) \tag{40}\]
(40) shows that the $\epsilon$-greedy policy is improvement. Furthermore, (37) can be represented as

$$V_\pi(s) = \sum_a \pi(a|s)Q_\pi(s, a) + (1 - \epsilon) \max_a Q_\pi(s, a)$$

$$= \frac{\epsilon}{A[s]} \sum_a Q_\pi(s, a) + (1 - \epsilon) \max_a Q_\pi(s, a)$$

$$= \frac{\epsilon}{A[s]} \sum_a \sum_{s', r} Pr_t(s', r|s, a)[r + \alpha V_\pi(s')]$$

$$+(1 - \epsilon) \max_a \sum_{s', r} Pr_t(s', r|s, a)[r + \alpha V_\pi(s')]$$  (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 TowerESCO Market. Accessed: Dec. 2015. [Online]. Available: http://www.gsmama.com/mobilefordevelopment/wp-content/uploads/2015/01/140617-GSAMAA- 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] Y. Chen, S. Zhang, S. Xu, and G. Li, “Fundamental trade-offs on green wireless networks,” IEEE Commun. Mag., vol. 49, no. 6, pp. 30–37, Jun. 2011.

[5] A. Abrol and R. K. Jha, “Power optimization in 5G networks: A step towards Green communication,” IEEE Commun. Mag., vol. 4, pp. 1355–1374, 2016.

[6] R. Mahapatra, Y. Nisjure, G. Kaddoum, N. Ull Hassan, and C. Yuen, “Energy efficiency tradeoff mechanism towards wireless green communication: A survey,” IEEE Commun. Surveys Tuts., vol. 18, no. 1, pp. 686–705, 1st Quart., 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 (Marcus Windisch), “Study on energy efficient radio access network (EERAN) technologies,” Tech. Univ. Dresden, Dresden, Germany, Tech. Rep., 2009.

[9] Y. S. Soh, T. Q. S. 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, J. Gong, Z. Yang, Z. Niu, and P. Yang, “Green mobile access network with dynamic base station energy saving,” in Proc. ACM MobiCom, 2009, vol. 9, no. 262, pp. 10–12.

[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 Qual. Service (IWQoS), Boston, MA, USA, Jun. 2010, pp. 1–2.

[12] D. Wilkomm, S. Machiraju, J. Bolot, and A. Wolszcz, “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. IIFIP Wireless Days (WD), Valencia, Spain, Nov. 2013, pp. 1–6.

[14] R. Torrea-Duran, P. Tsafakakis, L. Vandendorpe, and M. Moonen, “A cell breathing approach in green heterogeneous networks,” in Proc. IEEE 15th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC), Toronto, ON, Canada, Jun. 2014, pp. 344–348.

[15] 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.

[16] M. Robat Mili, L. Musavian, K. A. 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.

[17] Q. Wu, W. Chen, M. Tao, J. Li, H. Tang, and J. Wu, “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.

[18] N. K. M. Madi, Z. B. M. Hanapi, M. Othman, and S. Subramaniam, “Link adaptive power control and allocation for energy-efficient downlink transmissions in LTE systems,” IEEE Access, vol. 6, pp. 18469–18483, 2018.

[19] E. E. Tsiropoulos, T. Kastrinogiannis, and S. Papavassiliou, “Uplink power control in QoS-aware multi-service CDMA wireless networks,” J. Commun., vol. 4, no. 9, pp. 654–668, Oct. 2009.

[20] T. Kastrinogiannis, E. E. Tsiropoulos, and S. Papavassiliou, “Utility-based uplink power control in CDMA wireless networks with real-time services,” in Proc. ADHOC NOW, Berlin, Germany, 2008, pp. 307–320.

[21] K. K. Nguyen, T. Q. Duong, N. A. Vien, N.-A. Le-Khac, and M.-N. Nguyen, “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.

[22] A. Deshmukh and R. Vaze, “Online energy-efficient packet scheduling for a common deadline with and without energy harvesting,” IEEE J. Sel. Areas Commun., vol. 34, no. 12, pp. 3661–3674, Dec. 2016.

[23] M. J. Neely, “Optimistic scheduling with worst case delay guarantees in single and multi-hop networks,” in Proc. IEEE INFOCOM, Shanghai, China, Apr. 2011, pp. 1728–1736.

[24] Y. Luo, P. Hong, and R. Su, “Energy-efficient scheduling and power allocation for energy harvesting-based D2D communication,” in Proc. GLOBECOM-IEEE Global Commun. Conf., Singapore, Dec. 2017, pp. 1–6.

[25] W. Huang, W. Chen, and H. V. Poor, “Energy efficient wireless pushing with request delay information and delivery delay constraint,” IEEE Access, vol. 5, pp. 15428–15441, 2017.

[26] Y. Cui, V. K. N. 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] 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 Things J., vol. 6, no. 3, pp. 5585–5594, Jun. 2019.

[28] R. Sutton and A. Barto, Reinforcement Learning: An Introduction. Cambridge, MA, USA: MIT Press, Jan. 2018.

[29] W. Powell, Approximate Dynamic Programming: Solving the Curses of Dimensionality (Wiley Series in Probability and Statistics), vol. 24, no. 1.

[30] Y. Cui, V. K. N. 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.

[31] T. Kastrinogiannis, E. E. Tsiropoulos, and S. Papavassiliou, “Utility-based uplink power control in CDMA wireless networks with real-time services,” in Proc. ADHOC NOW, Berlin, Germany, 2008, pp. 307–320.

[32] R. Sutton and A. Barto, Reinforcement Learning: An Introduction. Cambridge, MA, USA: MIT Press, Jan. 2018.

[33] W. Powell, Approximate Dynamic Programming: Solving the Curses of Dimensionality (Wiley Series in Probability and Statistics), vol. 24, no. 1, 2007, p. 155.

[34] W. Powell, Approximate Dynamic Programming. Hoboken, NJ, USA: Wiley, 2011.

[35] I. Ahmed, K. T. Phan, and T. Le-Ngoc, “Optimal stochastic power control for energy harvesting wireless communication systems with a common deadline with and without energy harvesting,” IEEE J. Sel. Areas Commun., vol. 34, no. 12, pp. 3512–3527, Dec. 2016.

[36] C. Singhal, D. Swades, Resource Allocation in Next-Generation Broadband Wireless Access Networks. Hershey, PA, USA: IGI Global, 2017.
GUO JIANXIN received the B.S. degree in communications engineering from the Telecommunications Engineering Institute, Air Force Engineering University, Xi’an, China, in 1997, the M.S. degree in information and communications engineering from Air Force Engineering University, in 2000, and the Ph.D. degree in information and communications engineering from PLA Information Engineering University, Zhengzhou, China, in 2004.

From 2005 to 2007, he was a Lecturer with the Telecommunication Engineering Institute, Air Force Engineering University, where he was an Assistant Professor, from 2008 to 2009. 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 and more than 30 articles. His research interests include green communication, cognitive radio networks, and wireless resource allocation.

WANG FENG received the B.S. degree in communications engineering from Air Force Engineering University (AFEU), Xi’an, China, in 2001, and the Ph.D. degree in information and communications engineering from Xidian University Xi’an, in 2016.

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

BAOQIN LIN (Member, IEEE) received the M.Sc. degree in electromagnetic field and microwave technology from Air Force Engineering University, Xi’an, China, in 2002, and the Ph.D. degree in electronic science and technology from the National University of Defense Technology, Changsha, China, in 2006.

From 2002 to 2006, he was a Research Assistant with the Telecommunication Engineering Institute, Air Force Engineering University, where he was a Lecturer, from 2007 to 2016. Since 2017, he has been an Assistant Professor with the Information Engineering School, Xijing University, where he is currently an Associate Professor with the Department of Information Engineering. 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 the IEEE Antennas and Propagation Society.

YANGCHAO HUANG received the B.S. degree in communications engineering from the Telecommunications Engineering Institute, Air Force Engineering University, Xi’an, China, in 2001, the M.S. degree in information and communications engineering from Air Force Engineering University, in 2004, and the Ph.D. degree in information and communications engineering from PLA Information Engineering University, Xi’an, in 2016.

From 2004 to 2006, he was a Research Assistant with the Telecommunication Engineering Institute, Air Force Engineering University, where he was a Lecturer, from 2007 to 2016. Since 2017, he has been an Assistant Professor with the 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 networks, and communication counter.