Sensing, Probing, and Transmitting Policy for Energy Harvesting Cognitive Radio With Two-Stage After-State Reinforcement Learning

Keyu Wu, Hai Jiang, *Senior Member, IEEE*, and Chintha Tellambura, *Fellow, IEEE*

***Abstract*—This paper considers joint optimization of spectrum sensing, channel probing, and transmission power control for a single-channel secondary transmitter that operates with harvested energy from ambient sources. At each time slot, to maximize the expected secondary throughput, the transmitter needs to decide whether or not to perform the operations of spectrum sensing, channel probing, and transmission, according to energy status and channel fading status. First, we model this stochastic optimization problem as a two-stage continuous-state Markov decision process, with a sensing-and-probing stage and a transmit-power-control stage. We simplify this problem by a more useful after-state value function formulation. We then propose a reinforcement learning algorithm to learn the after-state value function from data samples when the statistical distributions of harvested energy and channel fading are unknown. Numerical results demonstrate learning characteristics and performance of the proposed algorithm.**

***Index Terms*—Cognitive radio, energy harvesting, power control, reinforcement learning, spectrum sensing.**

I. Introduction

Energy harvesting and cognitive radio aim to improve energy and spectral efficiency, respectively, of wireless networks. Wireless energy harvesting may prolong the battery lifetime of a wireless node, paving the way to greener communications [2]. Cognitive radio relieves the problems of scarcity and underutilization of spectrum [3]. Specifically, although the spectrum has been more or less fully allocated, temporarily unused spectrum slots of licensed or primary users (PUs) at specific locations result in spectrum holes. Therefore, unlicensed users (also called cognitive or secondary users [SUs]) sense the environment, detect spectrum holes, and opportunistically access the spectrum holes for their data transmission. Thus, one can get the best of both worlds by combining **energy harvesting and cognitive radio** [4]. However, the randomness of the energy harvesting process and the uncertainty of spectrum holes introduce unique challenges in optimal design of such systems.

Specifically, rapid and reliable identification of spectrum holes is essential for cognitive radio. Furthermore, when accessing spectrum holes, an SU must adapt its transmit power depending on channel fading status, which is indicated by channel state information (CSI) [5], [6]. The CSI estimation process is referred to as channel probing: i.e., the SU transmits a pilot sequence (see [7]–[9] and references therein for pilot designs), which enables its receiver to evaluate the channel and provide CSI feedback. Note that this channel probing should take place on an identified spectrum hole. But due to spectrum sensing errors, the SU may mistakenly estimate a channel (which is actually busy) to be available. This generates interference on PUs during both the channel probing and secondary data transmission stages. Thus, SU must minimize interference on PUs during all of its operations including spectrum sensing, channel probing and data transmission.

An energy harvesting SU may not perform all operations described above. For instance, if the PU channel is very likely to be occupied, the SU may skip sensing to save energy. In a deep fading channel, the SU may skip data transmission. Furthermore, since these three operations consume the harvested energy, they are coupled. Thus, in energy harvesting cognitive radio, it is important to jointly control the processes of sensing, probing and transmitting, by considering fading status, PU channel occupancy, and energy status.

A. Related Works

Sensing and/or transmission policies for energy harvesting cognitive radios have been extensively investigated [10]–[22], which are categorized and summarized below.

1) Optimal Sensing Design: Optimal sensing is investigated in [10]–[14] (without optimizing data transmission). Sensing policy (i.e., to sense or not) and energy detection threshold are derived for single-channel systems under an energy causality constraint in [10], [11]. Specifically, in [10], for a static (non-fading) channel, optimal sensing policy and energy detection threshold are derived by using the tool of Markov decision process (MDP), taking a collision constraint into account. In [11], sensing duration and energy detection threshold over a static channel are jointly optimized by using an MDP for a greedy sensing policy. Reference [12] considers multi-user multi-channel systems where the SUs harvest energy from PU signals. Balancing the goals of harvesting more energy (from busy channels) and gaining more access opportunities (from idle channels), the optimal SU scheduling problem (which schedules SUs to sense different channels) is investigated over fading channels, by using decentralized learning. In cooperative spectrum

Manuscript received May 3, 2018; revised October 4, 2018 and December 10, 2018; accepted December 15, 2018. Date of publication December 20, 2018; date of current version February 12, 2019. This work was supported by the Natural Sciences and Engineering Research Council of Canada. This paper was presented in part at the 2017 IEEE International Conference on Communications, Paris, France, May 2017 [1]. The review of this paper was coordinated by Dr. Y. Ma. *(Corresponding author: Keyu Wu.)*

The authors are with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 1H9, Canada (e-mail: keyu2@ualberta.ca; hai1@ualberta.ca; ct4@ualberta.ca).

Digital Object Identifier 10.1109/TVT.2018.2888826

sensing, the joint design of sensing policy, selection of cooperating SUs, and optimization of the sensing threshold is studied in [13] by using MDP. A similar problem is solved in [14] by using convex optimization, where the SUs harvest energy from both radio frequency and conventional (solar, wind and others) sources and different SUs have different sensing accuracy levels.

2) Optimal Transmission Control: This topic is considered in [15], [16]. Specifically, the work in [15] considers data rate adaptation and channel allocation for an energy-harvesting cognitive sensor node where channel availability status is provided by a third party (which does not deplete energy from the sensor node). Lyapunov optimization is used. Reference [16] uses convex optimization to jointly optimize time slot assignment and transmission power control in a time division multiple access system, assuming that the CSI between SUs and PUs is known. Here, the SUs use the underlay mode (i.e., they can transmit even if the PU spectrum is occupied, with a condition that the SUs' interference on PUs is not more than a certain threshold [23]).

3) Joint Optimization With Static (Non-Fading) Channels: Joint sensing and transmission design for static wireless channels is considered in [17]–[20]. Specifically, sensing policy, sensing duration and transmit power are jointly optimized by using an MDP in [17]. Similarly, sensing energy, sensing interval and transmit power are jointly designed by using an MDP in [18]. Reference [19] assumes an energy half-duplex constraint (i.e., sensing or transmitting is not allowed during energy harvesting). To balance energy harvesting, sensing accuracy and data throughput, a convex optimization method is used to jointly optimize the durations of harvesting, sensing, and transmission. Reference [20] considers that SUs harvest energy from PU signals. Durations of harvesting, sensing, and transmission are optimized by using convex optimization.

4) Joint Optimization With Fading Channels: The joint optimization of sensing and transmission with fading channels is studied in [21], [22]. Reference [21] investigates a heterogeneous secondary network that consists of energy-harvesting-powered spectrum sensors and battery-powered data sensors, which are jointly optimized (by using convex optimization) for maximizing overall energy efficiency and performance. Specifically, spectrum sensors are assigned over channels for maximizing the detected transmission opportunities. Given CSI of detected free channels, the data sensors determine the channels (to be used) and their transmission durations and transmission power levels over the channels, to minimize the overall energy consumption. Note that, in this work, the availability of CSI is assumed *a priori*. In [22], CSI acquisition is considered in a single-SU system. To this end, the SU probes CSI whenever energy is sufficient. Given probed CSI, the SU uses an MDP to decide which channel(s) to sense and whether to transmit or not if channel(s) are sensed free. Since the SU probes channels before spectrum sensing, the risk of probing busy channels exists. When this happens, the channel estimation pilots will be corrupted by PU signals, and the pilots may cause interference to primary receivers.

B. Motivations, Problem Statement and Contributions

Joint optimization of energy harvesting, channel sensing, probing, and transmission, especially over fading channels has not been reported widely. For instance, to adapt the transmit power according to fading status, channel probing is necessary, which can be conducted only if the PU channel is idle. Thus, the SU does not know its fading status when it decides whether or not to perform spectrum sensing. However, this *sensing-before-probing constraint* has not been captured before.

To fill this gap, we investigate a single-channel energy harvesting cognitive radio system. If the single channel is occupied by PUs, then the SU has no access. At each time slot, the SU decides whether to sense or not, and if the channel is sensed to be free, the SU may probe the channel. After a successful probing, the SU obtains CSI. With that, the SU needs to decide the transmit power level. To maximize long-term data throughput, we consider the joint optimization of sensing-probing-transmitting actions over a sequence of time slots.

In order to carry out optimal actions, the SU must track and exploit energy status, channel availability and fading status. These variables change randomly and are also affected by the previous sensing, probing and transmitting actions. We cast this stochastic dynamic problem under the framework of MDP [24] and reinforcement learning (RL) [25]. MDP is a mathematical tool for modeling stochastic optimal control. MDPs determine an optimal policy that maps each system state to an optimal action by considering the action's immediate reward and future effects. RL can solve the optimal policy of an MDP via exploiting samples collected from random rewards and state transitions. This is particularly useful when the exact model of the MDP is unknown or only partially known.

Although MDP and RL are standard tools, they should be carefully adapted for our problem. Specifically, due to the sensing-before-probing constraint, the SU cannot decide on its transmission power level (via adapting to CSI) when deciding whether or not to sense, since at the moment of making sensing decision, the SU has not obtained its CSI yet.

To incorporate the above feature in formulating and solving the optimal sensing-probing-transmitting policy, this paper makes the following contributions:

1) We devise a time-slotted protocol, where spectrum sensing, channel probing, and data transmission are conducted sequentially. We formulate the optimal decision problem as a two-stage MDP. The first stage deals with sensing and probing, while the second deals with the control of transmit power level. To the best of our knowledge, this is the first paper that separates the sensing-probing stage and the transmitting stage in MDP formulation for an energy harvesting SU.
2) Via exploiting the structure of the two-stage MDP, the optimal policy is developed based on an *after-state* (also called post-decision state) value function. The use of the after-state function confers three advantages. First, the solution of the original two-stage MDP presents practical and theoretical difficulties (Remarks 1 and 2 in Section III). The after-state value function can address these difficulties and derive the optimal policy (Remarks 3–5 and Corollary 1 in Section IV). Second, memory requirements to represent the optimal policy are minimized. Third, it enables the development of learning algorithms.
3) The SU often lacks the statistical distributions of harvested energy and channel fading. Thus, it must learn the optimal policy without this information. To achieve this, we propose an RL algorithm, which exploits samples of energy harvesting and channel fading in order to learn the after-state value function. The theoretical basis and performance bounds of the algorithm are also provided.

Notation convention: Meanings of important symbols are as follows. a: action; b: battery energy level; C: channel

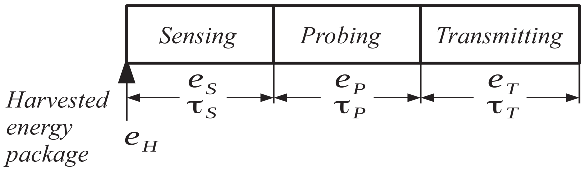


Fig. 1. Structure of a time slot.

availability status; d/x: endogenous/exogenous component of a state; $f_Y(\cdot)$: probability density function (pdf) of random variable (r.v.) Y; E_H/e_H: harvested energy (r.v./realization); H/h: channel gain (r.v./realization); $J^*(\cdot)$: after-state value function; p: belief; r: reward; s: state; Θ: sensing result; γ: discounting factor; β: after-state; ϵ: exploration rate. In subscripts and superscripts, t, S, P, T mean time slot index, "sensing", "probing", and "transmitting", respectively. A variable y' denotes the notation of y after one state transition in an MDP model.

II. System Model

We consider a single PU channel and one SU. The PU channel is shared by multiple PUs. All the PUs and the SU follow a time-slotted synchronous communication. Over the PU channel, the collective occupancy of the PUs across time slots is modeled as an on-off Markov process. States $C = 1$ and $C = 0$ denote that the PU channel is available and busy, respectively. The probability of state transition from state $i \in \{0, 1\}$ at a slot to state $j \in \{0, 1\}$ at the next slot is denoted as p_{ij}. It is assumed that the SU knows the state transition probability matrix, which can be estimated with long-term sensing measurements (see [26]). Note that the true state C is unknown by the SU. So the SU makes decisions based on all observed information (e.g., sensing results and others). All such information can be summarized as a scaled metric, known as the belief variable $p \in [0, 1]$, which represents the SU's belief in the channel's availability [27].

The SU always has data to send. A block fading model is applied. The channel gain between the SU and its receiver is H, which is an independent and identically distributed (i.i.d.) r.v. across time slots, with pdf $f_H(\cdot)$. This pdf is unknown to the SU.

The SU harvests energy from sources such as wind, solar, thermoelectric and others [28]. An energy package arrives at the beginning of each time slot (which was harvested throughout the previous time slot and stored in a temporal energy storage device [29], [30]). The energy amount E_H in the package is an i.i.d. r.v. across time slots, with pdf $f_E(\cdot)$. The SU does not know this pdf. The SU is equipped with a finite battery, with capacity $B_{\max}$. Let b denote the amount of energy stored in the SU's battery.

For the SU, each time slot is partitioned to three phases with τ_S, τ_P and τ_T for sensing, probing, and transmitting, respectively, shown in Fig. 1. In Fig. 1, e_H is the energy amount in the energy package that arrives at the beginning of the time slot, and e_S, e_P, e_T denote energy consumption in the three phases, respectively. Next we elaborate on the three phases of time slot t. A finite step machine (FSM) (see Fig. 2) is used to show the operations of the SU.

Sensing Step: At the beginning of the sensing phase of slot t, the SU, initially with battery level b_t^S, belief p_t^S, and harvested energy e_{Ht} (e_{Ht} means the amount of energy in the energy package that arrives at the beginning of slot t), needs to decide whether to sense or not. If the SU chooses not to sense (i.e., action $a_S = 0$ in transition (1.1) of Fig. 2), it remains idle until the beginning of slot $t + 1$ (i.e., the FSM transits to the sensing step of slot $t + 1$), at which time it has battery energy $b_{t+1}^S = \phi(b_t^S + e_{Ht})$, where $\phi(b)$ is defined as:

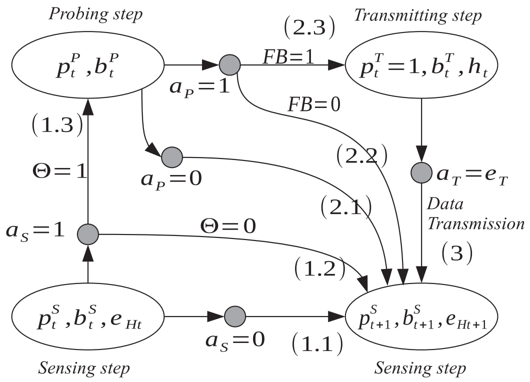


Fig. 2. FSM for sensing, probing, and transmitting of the SU.

$$\phi(b) \triangleq \max\{\min\{b, B_{\max}\}, 0\},$$

and the belief on channel availability changes to $p_{t+1}^S = \psi(p_t^S)$, where $\psi(p)$ is defined as:

$$\psi(p) \triangleq prob\{C_{t+1} = 1 | p_t = p\} = p \cdot p_{11} + (1 - p) \cdot p_{01},$$

which represents the SU's belief of next time slot given its belief of current time slot as p. Further, at the beginning of slot $t + 1$, an energy package arrives with energy amount e_{Ht+1}.

If the SU decides to sense at slot t (i.e., action $a_S = 1$), then during the sensing phase, it senses the PU channel, by using the energy detection method [31]. The sensing operation consumes a fixed amount of energy e_S. The sensing result is denoted as Θ: $\Theta = 0$ and $\Theta = 1$ mean that the SU estimates the PU channel to be busy and free, respectively. The performance of energy detector is characterized by a false alarm probability $p_{FA} \triangleq \Pr\{\Theta = 0 | C = 1\}$ and a miss-detection probability $p_M \triangleq \Pr\{\Theta = 1 | C = 0\}$. Here $\Pr\{\cdot\}$ means probability. Furthermore, $p_D \triangleq 1 - p_M$ and $p_O \triangleq 1 - p_{FA}$ represent the probability of correct detection of PU activities and the probability of a spectrum access opportunity, respectively. The values of p_{FA} and p_M are known to the SU.

We have the following observations for the sensing result.

1) The SU gets a negative sensing result (i.e., $\Theta = 0$) with probability $1 - G_1(p_t^S)$ (see transition (1.2) of Fig. 2), where $G_1(p)$ represents the probability of getting sensing result $\Theta = 1$ given initial belief p, i.e.,

$$G_1(p) \triangleq \Pr\{\Theta = 1 | p\} = p \cdot p_O + (1 - p) \cdot p_M.$$

Then the SU will remain idle until the beginning of slot $t + 1$, and we have $b_{t+1}^S = \phi(\phi(b_t^S + e_{Ht}) - e_S)$, and $p_{t+1}^S = \psi(G_2(p_t^S))$, where $G_2(p)$ means the probability that the channel is indeed idle given initial belief p and negative sensing result, i.e.,

$$G_2(p) \triangleq \Pr\{C = 1 | p, \Theta = 0\} = \frac{p \cdot p_{FA}}{p \cdot p_{FA} + (1 - p) \cdot p_D}.$$

2) The SU gets a positive sensing result ($\Theta = 1$) with probability $G_1(p_t^S)$ (see transition (1.3) of Fig. 2). Then the SU proceeds to the probing phase. At the beginning of the

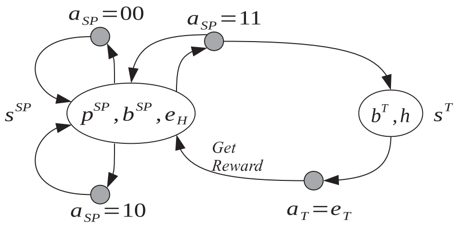


Fig. 3. Two-stage MDP.

probing phase, the battery level is $b_t^P = \phi(\phi(b_t^S + e_{Ht}) - e_S)$, and the belief transits to $p_t^P = G_3(p_t^S)$, where $G_3(p)$ is the probability that the channel is indeed idle, given initial belief p and positive sensing result, i.e.,

$$G_3(p) = \Pr\{C = 1 | p, \Theta = 1\} = \frac{p \cdot p_O}{p \cdot p_O + (1 - p) \cdot p_M}.$$

Probing Step: At the beginning of the probing phase, with information (p_t^P, b_t^P), the SU decides whether or not to probe the channel.[1] If it decides not to probe (i.e., action $a_P = 0$, see transition (2.1) of Fig. 2), the SU keeps inactive until the beginning of slot $t + 1$ (i.e., FSM transits to sensing step of slot $t + 1$), and the battery level remains the same $b_{t+1}^S = b_t^P$, and the belief becomes $p_{t+1}^S = \psi(p_t^P)$. If the SU decides to probe (i.e., action $a_P = 1$), it transmits channel estimation pilots (with energy consumption e_P) to the receiver.

1) There is probability $(1 - p_t^P)$ that the channel at slot t is busy (see transition (2.2) of Fig. 2). If this happens, the pilots will collide with primary activities, and will not be correctly received by the receiver. Thus, there will be no feedback (FB) from the receiver, denoted as $FB = 0$. Then the SU remains idle until the beginning of slot $t + 1$ with battery $b_{t+1}^S = \phi(b_t^P - e_P)$ and belief $p_{t+1}^S = p_{01}$.

2) There is probability p_t^P that the channel at slot t is idle, and the SU can get FB (denoted as $FB = 1$) and obtain the channel gain information, $h_t \geq 0$ (see transition (2.3) of Fig. 2). The SU then proceeds to the transmitting step. At this moment, the SU knows that the PU channel is free, i.e., $p_t^T = 1$, and the remaining energy is $b_t^T = \phi(b_t^P - e_P)$.

Transmitting Step: In transmitting step, action a_T is the amount of energy e_T to use for transmission during the transmitting phase. e_T is selected from a finite set $\mathbf{E_T}$ of energy levels. Note that if $e_T = 0$, there will be no transmission. After data transmission, it goes to the beginning of slot $t + 1$ with battery level $b_{t+1}^S = \phi(b_t^T - e_T)$ and belief $p_{t+1}^S = p_{11}$ (see transition (3) of Fig. 2).

III. Two-Stage MDP Formulation

A. Two-Stage MDP

Based on the FSM, we will use an MDP, shown in Fig. 3, to model the control problem. With s denoting a "state", a denoting an "action", an MDP is fully characterized by specifying the 4-tuple ($\mathbb{S}$, $\{\mathbb{A}(s)\}_s$, $f(\cdot|s, a)$, $r(s, a)$), namely state space, allowed actions at different states, state transition kernel, and reward associated with each state-action pair, which are described as follows.

To reduce the state space, we merge the sensing and probing steps into one stage (superscript SP) via jointly deciding these actions at the beginning of the sensing phase. We also observe that, at the transmitting step, the belief is always equal to 1, and thus, it is not necessary to represent it. Therefore, the state space $\mathbb{S}$ is divided into two classes: 1) sensing-probing state $s^{SP} = [b^{SP}, p^{SP}, e_H]$, with $b^{SP} \in [0, B_{\max}]$, $p^{SP} \in [0, 1]$ and $e_H \in [0, \infty)$; and 2) transmitting state $s^T = [b^T, h]$, with $b^T \in [0, B_{\max}]$ and channel gain $h \in [0, \infty)$. Note that, physically, b^{SP} and p^{SP} denote the battery level and belief value, respectively, at the beginning of a sensing phase. That is, at slot t, $b_t^{SP} = b_t^S$ and $p_t^{SP} = p_t^S$.

At a sensing-probing state s^{SP}, the full set of available actions are "not to sense" (action '00'), "to sense but not to probe" (action '10'), and "to sense and to probe if possible" (action '11'). Here for action 'yz', 'y' and 'z' mean the sensing decision and probing decision, respectively. So, we have $a_{SP} \in \mathbb{A}(s^{SP}) = \{00, 10, 11\}$. At a transmitting state s^T, the available actions are "transmission energy levels to use", i.e., $a_T \in \mathbb{A}(s^T) = \mathbf{E_T}$. As shown in Fig. 3, from a sensing-probing state, action '00' and '10' make a transition to a sensing-probing state (in next slot), while action '11' makes a transition to a transmitting state if the channel is sensed free and $FB = 1$, or to a sensing-probing state (in next slot) otherwise. From a transmitting state, it always transits to a sensing-probing state in the next slot.

$f(\cdot|s, a)$ is the pdf of the next state s' over $\mathbb{S}$ given initial state s and the taken action a. Denote $\delta(\cdot)$ as the Dirac delta function, which is used to generalize $f(\cdot|s, a)$ to include discrete transition components. We can derive the state transition kernel following the description of the FSM. Starting from $s_t^{SP} = [p_t^{SP}, b_t^{SP}, e_{Ht}]$, it may transit to $s_{t+1}^{SP} = [p_{t+1}^{SP}, b_{t+1}^{SP}, e_{Ht+1}]$ or $s_t^T = [b_t^T, h_t]$ depending on chosen actions[2], with $f(\cdot|s_t^{SP}, a_{SP})$ shown in (3), (4), (5) and (6) (on the bottom of next page). From transmitting state $s_t^T = [b_t^T, h_t]$, it can only transit to $s_{t+1}^{SP} = [p_{t+1}^{SP}, b_{t+1}^{SP}, e_{Ht+1}]$, with $f(\cdot|s_t^T, a_T)$ shown in (7) (on the bottom of next page). Note that we treat $f_H(\cdot)$ and $f_E(\cdot)$ as generalized pdf's, which cover discrete or mixed r.v. models for H and E_H.

At a sensing-probing state, because no data transmission has occurred yet, the reward is set to 0, i.e.,

$$r\left(s_t^{SP}, a^{SP}\right) = 0. \tag{1}$$

At a transmitting state, the reward is the amount of transmitted data, which is modeled (via the Shannon formula) as

$$r(s_t^T, a_T = e_T) = \tau_T W \log_2\left(1 + \frac{e_T h_t}{\tau_T N_0 W}\right) \mathbf{1}(b_t^T \geq e_T), \tag{2}$$

where W is the PU channel bandwidth, N_0 is the thermal noise power spectrum density and $\mathbf{1}(\cdot)$ is an indicator function.

We next place a technical restriction on the r.v. H.

[1] When the available energy is low, the SU may select to sense in the sensing phase but not to probe in the probing phase. By sensing, the SU can update its belief about channel availability, which can benefit its future decisions. As the available energy is low (e.g., the energy is insufficient to support a transmission), the SU may select not to probe, to save energy.

[2] During the transition from state $s_t^{SP} = [p_t^{SP}, b_t^{SP}, e_{Ht}]$, if probing is not carried out, the battery level and belief value are updated once; if probing is carried out, the battery level and belief are updated two times. For example, if action is '11' and the channel is sensed to be free ($\Theta = 1$) (which means the SU will probe) and the channel is indeed free ($C = 1$), then 1) b_t^{SP} first becomes $A = \phi(\phi(b_t^{SP} + e_{Ht}) - e_S)$ due to sensing operation, and becomes $\phi(A - e_P)$ due to probing operation, leading to $b_t^T = \phi(\phi(\phi(b_t^{SP} + e_{Ht}) - e_S) - e_P)$, and 2) p_t^{SP} first becomes $G_3(p_t^{SP})$ due to sensing operation, and becomes 1 due to probing operation, leading to $p_t^T = 1$.

Assumption 1: Given any battery level b^T and any transmission energy e_T, $\mathbb{E}[r(s^T, e_T)]$ and $\mathbb{E}[r^2(s^T, e_T)]$ exist and are bounded by some constants L_1 and L_2, respectively, with $\mathbb{E}[\cdot]$ being the expectation operation over r.v. H.

Comparing with one-stage MDP: Here, we clarify the difference between our proposed two-stage MDP and the one-stage MDPs of [17], [18], [22]. In these one-stage MDPs, states are defined as available information before performing spectrum sensing, where the sensing and transmission decision are made simultaneously. This is possible, as CSI h is assumed to be available before sensing the channel in [17], [18], [22]. Specifically, the works in [17], [18] assume a static channel (i.e., $h = 1$); while the work in [22] performs channel probing before spectrum sensing, which is, however, an unusual order.

In our problem, due to the sensing-before-probing constraint, one-stage MDP does not apply, and we need to divide the state space into two subspaces, one for sensing-probing decision making, and the other for transmission decision making, i.e., a two-stage MDP. This formulation naturally tracks and represents information-decision flow both *across* time slots (from s_t^{SP} to s_{t+1}^{SP}) and *within* a time slot (from s_t^{SP} to s_t^T). It enables us to apply generic MDP theory (Section III-B) to define the optimal policy. In addition, the solving of the optimal policy via after-state technique (Section IV) and RL algorithm (Section V) relies on analyzing the structure of the two-stage model.

B. Optimal Control via State Value Function V^*

Let Π denote all stationary deterministic policies, which are mappings from $s \in \mathbb{S}$ to $\mathbb{A}(s)$. We limit the control within Π. For any $\pi(\cdot) \in \Pi$, we define a function $V^\pi(\cdot) : \mathbb{S} \to \mathbb{R}$ for $\pi(\cdot)$ as follows,

$$V^\pi(s) \triangleq \mathbb{E}\left[\sum_{\tau=0}^{\infty} \gamma^\tau r(s_\tau, \pi(s_\tau)) | s_0 = s\right], \tag{8}$$

where s_τ denotes the state of time τ, $\gamma \in [0, 1)$ is a constant known as discounting factor,[3] and the expectation is defined[4] by the state transition kernel (3)–(7), shown at the bottom of the page. Therefore, by setting γ to a value that is close to 1, $V^\pi(s)$ can be (approximately) interpreted as the expected data throughput achieved by policy $\pi(\cdot)$ over infinite time horizon with initial state s.

Among Π, there is an optimal policy $\pi^*(\cdot) \in \Pi$ such that $V^{\pi^*}(s) = \sup_{\pi(\cdot)\in\Pi}\{V^\pi(s)\}, \forall s$, i.e., $\pi^*(\cdot)$ is able to maximize expected throughput for any initial state. In addition, $\pi^*(\cdot)$ can be identified by the Bellman equation [24, p. 154], which is defined as follows,

$$V(s) = \max_{a\in\mathbb{A}(s)} \{r(s, a) + \gamma\mathbb{E}[V(s')|s, a]\}, \tag{9}$$

where s' means the random next state given current state s and the taken action a. The state value function $V^*(s)$ is the solution to (9). Given $V^*(s)$, the optimal policy $\pi^*(s)$ can be defined as

$$\pi^*(s) = \arg\max_{a\in\mathbb{A}(s)} \{r(s, a) + \gamma\mathbb{E}[V^*(s')|s, a]\}. \tag{10}$$

Furthermore, it is shown [24, p. 152] that

$$V^*(s) = V^{\pi^*}(s), \quad \forall s. \tag{11}$$

Therefore, $V^*(s)$ and $V^{\pi^*}(s)$ are used interchangeably.

Remark 1: Although the optimal policy $\pi^*(s)$ can be obtained from the state value function $V^*(s)$, there are two practical difficulties for using (9) and (10) to solve our problem. First, the SU does not know the pdf's $f_E(\cdot)$ and $f_H(\cdot)$. The $\max\{\cdot\}$ operation in (9), which is performed over the $\mathbb{E}[\cdot]$ operation, makes it difficult to estimate[5] $V^*(s)$ by using samples. Second, $\mathbb{E}[\cdot]$ operation in (10) makes it difficult to get the optimal action, even if $V^*(s)$ is known.

Remark 2: In addition, there is another theoretical difficulty. In discounting MDP theory, the existence of $V^*(s)$ is usually established from the contraction theory, which requires the reward function $r(s, a)$ to be bounded for all s and all a [24, p. 143]. However, this is not satisfied in our problem, since we allow the channel gain h to take all positive values, and hence, $r(s, a)$ is unbounded over the state space. Therefore, in this case, the existence of $V^*(s)$ is not easy to establish.

As we will show in Section IV, both the practical and theoretical difficulties can be solved by transforming the value function into an after-state setting. Moreover, this transformation reduces space complexity via eliminating the explicit need for representing E_H and H processes.

IV. After-State Reformulation

Here, Section IV-A first analyzes the structure of the two-stage MDP. Then Section IV-B reformulates the optimal control

$$f\left(s_{t+1}^{SP}|s_t^{SP}, a_{SP} = 00\right) = \delta(p_{t+1}^{SP} - \psi(p_t^{SP}))\delta(b_{t+1}^{SP} - \phi(b_t^{SP} + e_{Ht}))f_E(e_{Ht+1}), \tag{3}$$

$$\begin{aligned} f(s_{t+1}^{SP}|s_t^{SP}, a_{SP} = 10) = {} & [(1 - G_1(p_t^{SP}))\delta(p_{t+1}^{SP} - \psi(G_2(p_t^{SP}))) + G_1(p_t^{SP})\delta(p_{t+1}^{SP} - \psi(G_3(p_t^{SP})))] \\ & \times \delta(b_{t+1}^{SP} - \phi(\phi(b_t^{SP} + e_{Ht}) - e_S))f_E(e_{Ht+1}), \end{aligned} \tag{4}$$

$$\begin{aligned} f(s_{t+1}^{SP}|s_t^{SP}, a_{SP} = 11) = {} & G_1(p_t^{SP})(1 - G_3(p_t^{SP}))\delta(p_{t+1}^{SP} - p_{01})\delta(b_{t+1}^{SP} - \phi(\phi(b_t^{SP} + e_{Ht}) - e_S - e_P)) \\ & \times f_E(e_{Ht+1}) + (1 - G_1(p_t^{SP}))\delta(p_{t+1}^{SP} - \psi(G_2(p_t^{SP})))\delta(b_{t+1}^{SP} - \phi(\phi(b_t^{SP} + e_{Ht}) - e_S))f_E(e_{Ht+1}), \end{aligned} \tag{5}$$

$$f(s_t^T|s_t^{SP}, a_{SP} = 11) = G_1(p_t^{SP})G_3(p_t^{SP})\delta(b_t^T - \phi(\phi(b_t^{SP} + e_{Ht}) - e_S - e_P))f_H(h_t). \tag{6}$$

$$f(s_{t+1}^{SP}|s_t^T, a_T = e_T) = \delta(p_{t+1}^{SP} - p_{11})\delta(b_{t+1}^{SP} - \phi(b_t^T - e_T))f_E(e_{Ht+1}). \tag{7}$$

[3] The discounting factor is used to ensure the infinite summation in (8) is bounded, and therefore, $V^\pi(s)$ is well defined.

[4] The expectation is taken over the random states $\{s_\tau\}_{\tau=1}^\infty$ with the distribution of s_τ determined by $f(\cdot|s_{\tau-1}, \pi(s_{\tau-1}))$.

[5] This difficulty can be illustrated with a simpler task. Given V^1 and V^2 are two r.v.s, suppose that we wish to estimate $\max\{\mathbb{E}[V^1], \mathbb{E}[V^2]\}$. And we can only observe a batch of samples $\{\max\{v_i^1, v_i^2\}\}_{i=1}^L$, where v_i^1 and v_i^2 are realizations of V^1 and V^2, respectively. However, the simple sample average of the observed information is not able to provide an unbiased estimation of $\max\{\mathbb{E}[V^1], \mathbb{E}[V^2]\}$, since $\lim_{L\to\infty} \frac{1}{L}\sum_{i=1}^L \max\{v_i^1, v_i^2\} \geq \max\{\mathbb{E}[V^1], \mathbb{E}[V^2]\}$.

TABLE I
STRUCTURED STATE TRANSITION MODEL

	d	x	$a \in \mathbb{A}(d,x)$	$\mathcal{N}(a)$	Observation	$p_i(d,a)$	$d' = \varrho_i([d,x],a)$	$\mathrm{f}_X(x'\|\varrho_i)$
s^{SP}	$[b,p]$	e_H	00	1	none	1	$[\psi(p), \phi(b+e_H)]$	$\mathrm{f}_E(\cdot)$
			10	2	$\Theta = 1$	$G_1(p)$	$[\psi(G_3(p)), \phi(\phi(b+e_H)-e_S)]$	$\mathrm{f}_E(\cdot)$
					$\Theta = 0$	$1 - G_1(p)$	$[\psi(G_2(p)), \phi(\phi(b+e_H)-e_S)]$	$\mathrm{f}_E(\cdot)$
			11	3	$\Theta = 1$, FB =1	$G_1(p)G_3(p)$	$\phi(\phi(b+e_H)-e_S-e_P)$	$\mathrm{f}_H(\cdot)$
					$\Theta = 1$, FB =0	$G_1(p)(1-G_3(p))$	$[p_{01}, \phi(\phi(b+e_H)-e_S-e_P)]$	$\mathrm{f}_E(\cdot)$
					$\Theta = 0$	$1 - G_1(p)$	$[\psi(G_2(p)), \phi(\phi(b+e_H)-e_S)]$	$\mathrm{f}_E(\cdot)$
s^T	b	h	e_T	1	none	1	$[p_{11}, \phi(b-e_T)]$	$\mathrm{f}_E(\cdot)$

in terms of after-state value function J^*. Finally, the solution of J^*, and its relationships with the state value function V^* are given in Section IV-C.

A. Structure of the MDP

The structural properties of the MDP given in the 4-tuple $(\mathbb{S}, (\mathbb{A}(s))_s, f(\cdot|s,a), r(s,a))$ are as follows.

1) We divide each state into endogenous and exogenous components. Specifically, for a sensing-probing state s^{SP}, the endogenous and exogenous components are $d^{SP} = [p^{SP}, b^{SP}]$ and $x^{SP} = \{e_H\}$, respectively. All possible d^{SP} and x^{SP} are defined as $\mathbb{D}^{SP}$ and $\mathbb{X}^{SP}$, respectively. Similarly, for a transmitting state s^T, the endogenous and exogenous components are $d^T = \{b^T\}$ and $x^T = \{h\}$, respectively. All possible d^T and x^T are $\mathbb{D}^T$ and $\mathbb{X}^T$, respectively.
   Finally, let $d \in \mathbb{D} = \mathbb{D}^{SP} \cup \mathbb{D}^T$ and $x \in \mathbb{X} = \mathbb{X}^{SP} \cup \mathbb{X}^T$.
2) The number of available actions $\mathbb{A}(s)$ at each state s is finite.
3) Checking the state transition kernel (3), (4), (5), (6) and (7), we can see that, given state $s = [d,x]$, and action $a \in \mathbb{A}(s)$, the transition to next state $s' = [d', x']$ has following properties.
   - The stochastic model of d' is fully known. Specifically, after applying an action a taken at state $s = [d,x]$, we have $\mathcal{N}(a)$ possible cases depending on SU's observations (e.g., sensing-probing outcomes after applying certain a^{SP}). This leads to $\mathcal{N}(a)$ possible values of d'. And at the ith case, which happens with probability $p_i(d,a)$, the value of d' takes the value $\varrho_i(s,a)$. Functions $\mathcal{N}(\cdot)$, $\varrho_i(\cdot,\cdot)$ and $p_i(\cdot,\cdot)$ are known, and listed in Table I for different d, x, a and observations.
   - The x' is a r.v. whose distribution depends on $\varrho_i(s,a)$, i.e., if $\varrho_i(s,a) \in \mathbb{D}^{SP}$, x' has pdf $f_E(\cdot)$; and if $\varrho_i(s,a) \in \mathbb{D}^T$, x' has pdf $f_H(\cdot)$ (see Table I). This relationship is described by conditional pdf $f_X(x'|\varrho_i(s,a))$.

   With these notations, the state transition kernel $f(s'|s,a)$ can be rewritten as:
$$f(s'|s,a) = f((d',x')|(d,x),a)$$
$$= \sum_{i=1}^{\mathcal{N}(a)} p_i(d,a)\delta(d' - \varrho_i(s,a)) f_X(x'|\varrho_i(s,a)). \quad (12)$$
4) The reward $r([d,x],a)$ is deterministic, defined via (1) and (2).

B. Introducing After-State Based Control

Based on the above structural properties, we now show that optimal control can be developed based on so-called "after-states". Physically, an after-state is the endogenous component of a state. However, for ease of presentation, we consider it as a "virtual state" appended to the original MDP (Fig. 4).

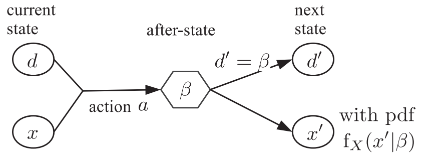


Fig. 4. Augmented MDP model with after-state.

Specifically, after an action a applied on a state $s = [d,x]$, the state randomly transits to an after-state β. The number of such transitions is $\mathcal{N}(a)$. At the ith transition, the after-state is $\beta = \varrho_i([d,x],a)$ with probability $p_i(d,a)$. From β, the next state is $s' = [d', x']$ with $d' = \beta$ and x' has pdf $f_X(\cdot|\beta)$.

We next introduce after-state based control. The main ideas are as follows. From β, the next state $s' = [d', x']$ only depends on β (i.e., $d' = \beta$, and the pdf of x' is conditioned on β). Therefore, ***starting from an after-state*** β***, the maximum expected discounted reward only depends on*** β, and is denoted as an *after-state value function* $J^*(\beta)$. The key is that if $J^*(\beta)$ is known for all β, the optimal action at a state $s = [d,x]$ can be determined as

$$\pi^*([d,x]) = \underset{a \in \mathbb{A}([d,x])}{\arg\max} \left\{ r([d,x],a) + \sum_{i=1}^{\mathcal{N}(a)} p_i(d,a) J^*(\varrho_i([d,x],a)) \right\}. \quad (13)$$

The expression in (13) is intuitive: the optimal action at a state $s = [d,x]$ is the one that maximizes the sum of the immediate reward $r([d,x],a)$ and the expected maximum future value $\sum_{i=1}^{\mathcal{N}(a)} p_i(d,a) J^*(\varrho_i([d,x],a))$. The solving of $J^*(\cdot)$ and the formal proof of (13) are provided in Section IV-C.

Remark 3: Unlike (10), if $J^*(\cdot)$ is known, generating actions with (13) is easy, since $\mathcal{N}(a)$ and $|\mathbb{A}(s)|$ are finite, and $p_i(d,a)$ and $\varrho_i([d,x],a)$ are known. Furthermore, the space complexity of $J^*(\cdot)$ is lower than that of $V^*(\cdot)$, since $\mathbb{X}$ does not need to be represented in $J^*(\cdot)$.

C. Establishing After-State Based Control

The development of this subsection is as follows. First, we define a so-called after-state Bellman equation as

$$J(\beta) = \gamma \mathop{\mathbb{E}}_{X'|\beta} \left[\max_{a' \in \mathbb{A}([\beta, X'])} \left\{ r(\beta, X', a') + \sum_{i=1}^{\mathcal{N}(a')} p_i(\beta, a') J(\varrho_i([\beta, X'], a')) \right\} \right], \quad (14)$$

where $\mathbb{E}_{X'|\beta}[\cdot]$ means taking expectation over r.v. X', which has pdf $f_X(\cdot|\beta)$. Note that X' means the random exogenous variable of the next state given that the current after-state is β (see Fig. 4). Then, Theorem 1 shows that (14) has a unique solution $J^*(\cdot)$, and also provides a value iteration algorithm for solving it. Note that, at this moment, the meaning of $J^*(\cdot)$ is unclear. Finally, Theorem 2 and Corollary 1 show that $J^*(\cdot)$ is exactly the after-state value function defined in Section IV-B, and the policy defined with (13) is equivalent with (10), and therefore, is the optimal policy.

Theorem 1: Given Assumption 1, there is a unique $J^*(\cdot)$ that satisfies (14). And $J^*(\cdot)$ can be calculated via a value iteration algorithm: with $J_0(\cdot)$ being an arbitrary bounded function, the sequence of functions $\{J_l(\cdot)\}_{l=0}^{L}$ defined by the following iteration equation: for all $\beta \in \mathbb{D}$,

$$J_{l+1}(\beta) \leftarrow \gamma \mathop{\mathbb{E}}_{X'|\beta}\left[\max_{a' \in \mathbb{A}([\beta, X'])}\left\{r([\beta, X'], a') + \sum_{i=1}^{\mathcal{N}(a')} p_i(\beta, a') J_l(\varrho_i([\beta, X'], a'))\right\}\right], \tag{15}$$

converges to $J^*(\cdot)$ when $L \to \infty$.

Proof: See Appendix A. ■

Remark 4: Unlike the classical Bellman equation (9), in the after-state Bellman equation (14), the expectation is outside of the reward function. While the reward function is unbounded, its expectation is bounded due to Assumption 1. Therefore, the solution to (14) can be established by contraction theory.

Remark 5: Comparing with (9), equation (14) exchanges the order of (conditional) expectation and maximization operators. And inside the maximization operator, functions $r(s, a)$, $\mathcal{N}(a)$, $p_i(d, a)$, and $\varrho_i(s, a)$ are known. These are crucial in developing a learning algorithm that uses samples to estimate the after-state value function $J^*(\cdot)$.

Theorem 2: The existence of a solution $V^*(s)$ to (9) can be established from $J^*(\beta)$. In addition, their relationships are

$$V^*([d, x]) = \max_{a \in \mathbb{A}([d,x])}\left\{r([d, x], a) + \sum_{i=1}^{\mathcal{N}(a)} p_i(d, a) J^*(\varrho_i([d, x], a))\right\} \tag{16}$$

and

$$J^*(\beta) = \gamma \mathop{\mathbb{E}}_{X'|\beta}\left[V^*([\beta, X'])\right]. \tag{17}$$

Proof: See Appendix B. ■

Corollary 1: $J^*(\cdot)$ is the after-state value function, and the policy defined with (13) is optimal.

Proof: From (17) and the physical meaning of $V^*(\cdot)$ (see (11)), $J^*(\beta)$ represents the maximum expected discounted sum of rewards, starting from after-state β. Therefore, $J^*(\cdot)$ is the after-state value function.

The expression in (13) can be derived from the optimal policy (10) as follows: first decompose the expectation with (12), and then plug in (17). Therefore, (13) is the optimal policy. ■

Corollary 1 shows that optimal control can be achieved equivalently through value function $J^*(\cdot)$. And Theorem 1 establishes the existence of $J^*(\cdot)$ and also provides a value iteration algorithm for solving $J^*(\cdot)$. To obtain $J^*(\cdot)$ using the value iteration algorithm, we have two observations. 1) The after-state space $\mathbb{D}$ is continuous. Thus, after-state space discretization is needed. 2) The computation of $\mathbb{E}_{X'|\beta}[\cdot]$ requires the knowledge of $f_E(\cdot)$ and $f_H(\cdot)$, which is unknown in our setting. Thus, RL can be used to learn a (near) optimal policy via sample averaging, instead of taking expectation. Details are given in the next section.

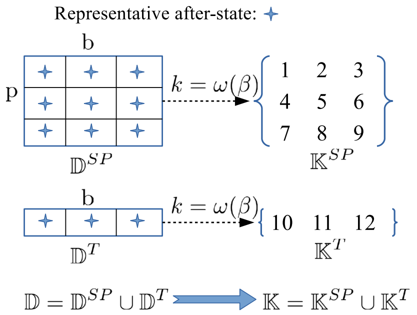


Fig. 5. An example of after-state space discretization.

V. Reinforcement Learning Algorithm

In this section, we first discretize the after-state space into finite clusters,[6] which is discussed in Section V-A. In addition, a learning algorithm is proposed in Section V-B, which learns a (near) optimal policy given data samples of energy harvesting and wireless channel. Furthermore, the algorithm's convergence guarantee and performance bounds are analyzed in Section V-C. Finally, the algorithm is modified in Section V-D, for achieving simultaneous data sampling, learning and control.

A. After-State Space Discretization

We divide the continuous after-state space $\mathbb{D}$ into a finite number of portions or clusters $\mathbb{K}$, which defines a mapping $\omega(\cdot): \mathbb{D} \to \mathbb{K}$. In addition, all after-states assigned into the same cluster are mapped into one representative after-state. Mathematically, let $\mathbb{D}(k) \triangleq \{\beta \in \mathbb{D} | \omega(\beta) = k\}$ denote the set of after-state assigned to cluster $k \in \mathbb{K}$. Thus, we use $q(k) \in \mathbb{D}(k)$ to represent all after-states of $\mathbb{D}(k)$.

As an example, in Fig. 5, two-dimensional $\mathbb{D}^{SP}$ is uniformly discretized into 9 clusters $\mathbb{K}^{SP} = \{1, \ldots, 9\}$. The one-dimensional after-state space $\mathbb{D}^{T}$ is uniformly discretized into 3 clusters $\mathbb{K}^{T} = \{10, 11, 12\}$. The association from an after-state β to the cluster k is denoted by $k = \omega(\beta)$. And the after-states assigned to the same cluster are represented by its central point, $q(k)$.

B. Learn Optimal Policy With Data Samples

With this discretization, we design an RL algorithm that learns near optimal policy from the samples of E_H and H.

The idea is to learn a function $g(k)$, for $k \in \mathbb{K}$, such that $g(\omega(\beta))$ is close to $J^*(\beta)$ for all $\beta \in \mathbb{D}$. Then a near-optimal

[6] In addition to discretization, other methods can also be used, such as tile coding or radial basis functions [25, Chapter 8.3], which may accelerate the learning.

policy can be constructed as

$$\begin{aligned}&\hat{\pi}([d,x]|g)\\ &= \arg\max_{a\in\mathbb{A}([d,x])} \left\{ r([d,x],a) + \sum_{i=1}^{\mathcal{N}(a)} p_i(d,a)\, g(\omega(\varrho_i([d,x],a))) \right\}.\end{aligned} \tag{18}$$

Comparing (18) with (13), we observe that if $g(\omega(\beta))$ approximates $J^*(\beta)$ accurately, $\hat{\pi}(\cdot|g)$ is close to $\pi^*(\cdot)$.

The function $g(k)$ is learned by iteratively updating with data samples. Each update uses only one data sample. This facilitates the tailoring of the algorithm for online applications (Section V-D). Next, we present the algorithm and some intuitive reasons.

1) Algorithm: Initially we have arbitrary bounded function $g_0(k)$. We calculate $g_{l+1}(k)$ from $g_l(k)$ and x_l, the lth data sample. Since x_l can be either an energy or fading sample, there are two cases:

- if x_l is a sample of E_H, randomly choose N non-repeated clusters from $\mathbb{K}^{SP}$;
- if x_l is a sample of H, randomly choose N non-repeated clusters from $\mathbb{K}^{T}$.

Here N is a parameter to balance learning speed and computation load. For either case, we denote the set of chosen clusters as $\bar{K}_l$. Given x_l and $\bar{K}_l$, we have the updating rule as

$$g_{l+1}(k) = \begin{cases} (1-\alpha_l(k))\cdot g_l(k) + \alpha_l(k)\cdot \delta_l(k), & \text{if } k \in \bar{K}_l;\\ g_l(k), & \text{otherwise,} \end{cases} \tag{19}$$

where $\alpha_l(k) \in (0,1)$ is the step size of cluster k for the lth iteration, and $\delta_l(k)$ is constructed with x_l as

$$\begin{aligned}\delta_l(k) \triangleq \gamma \max_{a\in\mathbb{A}([q(k),x_l])} &\Bigg\{ r([q(k),x_l],a)\\ &+ \sum_{i=1}^{\mathcal{N}(a)} p_i(q(k),a)\, g_l(\omega(\varrho_i([q(k),x_l],a))) \Bigg\}.\end{aligned} \tag{20}$$

Section V-C will show that with proper setting of the step size $\alpha_l(k)$, the sequence of functions $\{g_l(k)\}_{l=1}^{\infty}$ converges such that $g_\infty(\omega(\beta))$ is close to $J^*(\beta)$, and the policy $\hat{\pi}(\cdot|g_\infty)$ defined in (18) is close to $\pi^*(\cdot)$.

The above algorithmic pieces are summarized in Algorithm 1. For a sufficiently large number (L) of iterations, the learning process can be considered complete. The learned policy $\hat{\pi}(s|g_L)$ can then be used for sensing, probing and transmission control, just as in Algorithm 2 in Section V-D.

2) Intuition: Algorithm 1 is a stochastic approximation algorithm [32], which is intuitive generalization of the value iteration algorithm (15). Specifically, it is known from (15) that, given the value function $J_l(\cdot)$ of the l-th iteration, a noisy estimation of $J_{l+1}(\beta)$ can be constructed as

$$\max_{a'\in\mathbb{A}([\beta,x'])} \left\{ r([\beta,x'],a') + \sum_{i=1}^{\mathcal{N}(a')} p_i(\beta,a')\, J_l(\varrho_i([\beta,x'],a')) \right\}, \tag{21}$$

with x' sampled from $f_X(\cdot|\beta)$, i.e., x' is a realization of E_H if $\beta \in \mathbb{D}^{SP}$, and x' is a realization of H if $\beta \in \mathbb{D}^{T}$.

Algorithm 1: Learning of Control Policy.

```
Require: Data samples $\{x_l\}_l$
Ensure: Learned control policy $\hat{\pi}(s|g_L)$
  Initialize $g_0(k) = 0,\ \forall k$
  for $l$ from 0 to $L-1$ do
      if $x_l$ is a data sample of $E_H$ then
          Choose $N$ clusters from $\mathbb{K}^{SP}$ and get $\overline{K_l}$
      else if $x_l$ is a data sample of $H$ then
          Choose $N$ clusters from $\mathbb{K}^{T}$ and get $\overline{K_l}$
      end if
      Generate $g_{l+1}(\cdot)$ by executing (19) with ($x_l$, $\overline{K_l}$)
  end for
  With $g_L(\cdot)$, construct control policy $\hat{\pi}(\cdot|g_L)$ through (18)
```

Therefore, by comparing (21) with (20), we see $\delta_l(k)$ as an estimate of $g_{l+1}(k)$ for $k \in \bar{K}_l$ (with $\omega(\cdot)$ introduced for discretization, β approximated with $q(k)$, and $J_l(\cdot)$ replaced with $g_l(\cdot)$). Hence, with $\delta_l(k)$, equation (19) updates $g_{l+1}(\cdot)$ for chosen clusters within $\bar{K}_l$ by sample averaging. Note that, theoretically, we can set $\bar{K}_l$ to $\mathbb{K}^{SP}$ or $\mathbb{K}^{T}$ (x_l is energy or fading sample), which could accelerate learning speed. However, large $|\mathbb{K}^{SP}|$ or $|\mathbb{K}^{T}|$ leads to increased computations. Hence, the parameter N is introduced to control the computational burden. Section VI-B1 gives an example to show impact of N.

C. Theoretical Soundness and Performance Bounds

In this part, we formally state the convergence requirements and performance guarantees for Algorithm 1.

First, for $\forall\, k \in \mathbb{K}$, we define $M(k) = \{l \in \{0,1,\ldots,L-1\} | k \in \overline{K_l}\}$, which presents the set of iteration indices where k is chosen during learning. In addition, we define

$$\xi \triangleq \max_k \Big\{ \sup_{\beta\in\mathbb{D}(k)} |J^*(\beta) - J^*(q(k))| \Big\}, \tag{22}$$

which describes the "error" introduced by the after-state space discretization. Finally, in order to evaluate the performance of a policy $\pi(\cdot)$ from after-states' point of view, we define

$$J^{\pi}(\beta) = \gamma \mathop{\mathbb{E}}_{X'|\beta} \left[V^{\pi}([\beta, X']) \right], \tag{23}$$

where $V^{\pi}(\cdot)$ is defined in (8).

Given the definitions of $M(k)$, ξ and $J^{\pi}(\beta)$, we have following theorem.

Theorem 3: Given that Assumption 1 is true, and also assuming that, in Algorithm 1, as $L \to \infty$,

$$\sum_{l\in M(k)} \alpha_l(k) = \infty,\ \forall k \tag{24}$$

$$\sum_{l\in M(k)} \alpha_l^2(k) < \infty,\ \forall k \tag{25}$$

then we have:

i) the sequence of functions $\{g_l(\cdot)\}_{l=0}^{L}$ generated in (19) converge to a function $g_\infty(\cdot)$ with probability 1 as $L \to \infty$;

ii) $||J^* - J_\infty|| \le \frac{\xi}{1-\gamma}$, with

$$J_\infty(\beta) \triangleq g_\infty(\omega(\beta)), \tag{26}$$

and $||\cdot||$ denoting the maximum norm;

iii) $||J^* - J^{\pi_\infty}|| \le \frac{2\gamma\xi}{(1-\gamma)^2}$, with $\pi_\infty(\cdot) \triangleq \hat{\pi}(\cdot|g_\infty)$.

Proof: See Appendix C. ■

Remark 6: Statement (i) of Theorem 3 demonstrates the convergence guarantee of Algorithm 1. Statement (ii) shows that the learned function $g_\infty(\cdot)$ is close to $J^*(\cdot)$, and their difference is controlled by the error ξ caused by after-state space discretization. Statement (iii) claims that asymptomatically, the performance of policies $\{\hat{\pi}(\cdot|g_l)\}_l$ approaches that of the optimal policy $\pi^*(\cdot)$, and that the performance gap is proportional to the error ξ.

Remark 7: Condition (24) requires $|M(k)| = \infty$, where $|M(k)|$ denotes the size of $M(k)$. In other words, energy harvesting and wireless fading processes need to be sampled infinitely often in $\{x_l\}_{l=0}^{L-1}$, as $L \to \infty$.

In order to satisfy $\sum_{l \in M(k)} \alpha_l^2(k) < \infty$, the sequence of step size $\{\alpha_l(k)\}_{l \in M(k)}$ should start to decay after certain l with sufficient decay rate. However, the decay rate should not be too large, in order to satisfy $\sum_{l \in M(k)} \alpha_l(k) = \infty$.

D. Simultaneous Sampling, Learning and Control

Algorithm 1 operates offline — the policy is learned with given data samples, and a learned policy cannot be used until learning is complete. However, for some applications, an online learning scheme may be more desirable. In online case, sequential data is used to update the best learned policy at each step.

One intuitive idea to tailor Algorithm 1 for online learning is as follows. Supposing that current learned function is $g_l(\cdot)$, we can use $\hat{\pi}(\cdot|g_l)$ to generate actions and interact with the environment in real-time. Thus, we can collect a data sample from energy harvesting or channel fading, which can be further used to generate $g_{l+1}(\cdot)$. As the loop continues, $g_l(\cdot)$ approaches $g_\infty(\cdot)$, and the policy $\hat{\pi}(\cdot|g_l)$ approaches $\pi_\infty(\cdot)$, which implies that generated actions during the process will be more and more likely to be optimal. In this way, simultaneous sampling, learning and control can be achieved.

However, the problem is that the above method cannot guarantee to sample the wireless fading process infinitely-often (i.e., cannot satisfy assumptions (24) and (25) of Theorem 3). Note that the wireless fading process can be sampled only if $\hat{\pi}(\cdot|g_l)$ chooses $a^{SP} = 11$. But the above method may enter a deadlock such that $a^{SP} = 11$ will never be chosen. The deadlock can be caused by: (1) insufficient battery energy that results from the learned policy's consistent aggressive use of energy; and/or (2) persistently locking in $a^{SP} = 00$ or $a^{SP} = 10$. In order to break this possible deadlock during the learning process, with some small probability ϵ (named as the *exploration rate*), we force the algorithm to deviate from $\hat{\pi}(\cdot|g_l)$ to exploring the environment by either accumulating energy ($a^{SP} = 00$) or probing channel gain information ($a^{SP} = 11$).

Based on the above points, Algorithm 2 is provided for sampling, learning and control. Here, we argue that $g_l(\cdot)$ generated by Algorithm 2 converges to $g_\infty(\cdot)$ when $t \to \infty$. First, at each time slot, there is probability $\epsilon/2$ that the algorithm will choose $a^{SP} = 00$ to accumulate energy. Therefore, given the battery level b_t^{SP} of slot t, we can[7] find a finite T such that $prob\{b_{t+T}^{SP} \ge e_S + e_P\} > 0$. In other words, at any slot $t \ge T$, we have $prob\{b_t^{SP} \ge e_S + e_P\} > 0$. Thus, having sufficient energy for sensing and probing, the algorithm will choose $a^{SP} = 11$ with probability $\epsilon/2$. In addition, at any time slot, the PU channel will be free with a non-zero probability. Therefore, the algorithm can reach the transmitting stage with a non-zero probability. Thus, the wireless fading process can be sampled infinitely often for $t \to \infty$. In summary, the assumptions (24) and (25) of Theorem 3 are satisfied (under properly decayed step size), and $\{g_l(\cdot)\}_l$ converges to $g_\infty(\cdot)$ asymptotically.

Algorithm 2: Simultaneous Sampling, Learning and Control.

```
    Note: $\beta_t^{SP} \in \mathbb{D}^{SP}$ presents an after-state in slot $t$. $\beta_t^T$ is
    defined similarly.
 1: Initialize: battery $b_0$, channel belief $p_0$, and after-state
    $\beta_0^{SP} = [b_0, p_0]$
 2: Initialize: $g_0(k) = 0,\ \forall k$, and set $l = 0$
 3: for $t$ from 1 to $\infty$ do
 4:     Observe arriving harvested energy amount $e_{Ht}$
 5:     Set $x_l = e_{Ht}$ and choose $\overline{K}_l$ with $N$ clusters from
        $\mathbb{K}^{SP}$
 6:     Generate $g_{l+1}(\cdot)$ by executing (19) with $(x_l, \overline{K}_l)$
 7:     $l \leftarrow l + 1$
 8:     Construct state $s_t^{SP} = [\beta_{t-1}^{SP}, e_{Ht}]$
 9:     Generate sensing-probing decision $a_t^{SP} = \hat{\pi}(s_t^{SP}|g_l)$
        via (18)
10:     if random() $\le \epsilon$ then                    ▷ Exploration
11:         if random() $\le 1/2$ then
12:             $a_t^{SP} = 00$
13:         else if $b_t^{SP} \ge e_S + e_P$
            then                                  ▷ Energy sufficiency
14:             $a_t^{SP} = 11$
15:         end if
16:     end if
17:     Apply sensing and probing actions based on $a_t^{SP}$
18:     if $a_t^{SP} = 11$ & $\Theta = 1$ & $FB = 1$ then
19:         Observe the channel gain $h_t$ from FB
20:         Set $x_l = h_t$, and construct $\overline{K}_l$ by choosing $N$
            clusters from $\mathbb{K}^T$
21:         Generate $g_{l+1}(\cdot)$ by executing (19) with $(x_l, \overline{K}_l)$
22:         $l \leftarrow l + 1$
23:         Derive after-state $\beta_t^T$ with $s_t^{SP}$ via Table I
24:         Construct state $s_t^T = [\beta_t^T, h_t]$
25:         Generate transmit decision $a_t^T = \hat{\pi}(s_t^T|g_l)$
            via (18)
26:         Set transmission power based on $a_t^T$, and
            transmit data
27:         Derive after-state $\beta_t^{SP}$ from $(s_t^T, a_t^T)$ via Table I
28:     else
29:         Derive after-state $\beta_t^{SP}$ with $s_t^{SP}$ and $a_t^{SP}$, $\Theta$
            and $FB$ via Table I
30:     end if
31: end for
```

1) Complexity Analysis of Algorithm 2: For each t, major computations are the two embedded function updates for $g_l(\cdot)$ (line 6 and line 21). Each update needs to compute (20) N times. And each computation requires $|\mathcal{N}(a)|$ multiplications, $|\mathcal{N}(a)|$ summations, and one maximization over a set.

2) Choices of Exploration Rate: Although the convergence is guaranteed for any $\epsilon \in (0, 1)$, the choice of ϵ affects the performance of the algorithm. Large ϵ helps to accelerate the learning process. But too large ϵ may cause big loss of the achievable performance. See Section VI-B2 for examples.

[7] If this condition cannot be satisfied, the underlying energy harvesting process is not sufficient to power the SU.

VI. Simulation Results

A. Simulation Setup

We use simulation to evaluate the performance of the proposed algorithms. The simulation is set up as follows.

The PU channel occupancy Markov model is described by $p_{00} = 0.9$ and $p_{11} = 0.9$. For spectrum sensing, we have $p_{FA} = 0.2$ and $P_D = 0.9$. The time slot duration $\tau_S + \tau_P + \tau_T$ (Fig. 1) of the SU is 12 ms (which is synchronized to the PU channel). We set $\tau_S = \tau_P = 1$ ms.[8] Hence, within each time slot, we have $\tau_T = 10$ ms for data transmission.

Energy is harvested from wind power. Thus, E_H is well characterized by the Weibull distribution [33], with shape and mean parameters $k_E = 1.2$ and $\mu_E = 1$ (in Sections VI-C and VI-D, other values of μ_E are considered).

The SU signal channel gain h consists of path loss h_s and Rayleigh fading h_f. h_s is distance-dependent and is assumed to be fixed. h_f has pdf $f(x) = e^{-x}, \quad x \geq 0$.

Then, with above channel model, the amount of transmitted data can be rewritten as $\tau_T W \log_2(1 + \frac{e_T h_s h_f}{\tau_T N_0 W}) = \tau_T W \log_2(1 + \frac{e_T h_f}{\eta})$ where $\eta \triangleq \tau_T N_0 W / h_s$ with $W = 1$ MHz. We normalize η as $\eta = 1$ (for energy normalization). Normalizing with respect to η, we set battery capacity $B_{\max} = 10$, sensing energy $e_S = 1$, probing energy $e_P = 2$, and the set of transiting energy levels $\mathbf{E_T} = \{0, 3, 4, 5, 6\}$.

Finally, simple uniform grid is used for discretization with 10 levels for both belief and battery dimensions. Thus, $|\mathbb{K}^{SP}| = 100$ and $|\mathbb{K}^{T}| = 10$. We set the discounting factor γ as 0.9, and the learning step size rule as $\alpha_l(k) = \frac{10^4}{l + 10^4}$, where l is the index of updating iteration in Algorithm 1 and Algorithm 2.

B. Characteristics of Learning Algorithm

1) Offline Learning Under Various N: Here, we study the learning speed of our offline algorithm (Algorithm 1) under different N. We set the updating iteration budget $L = 10^6$, and x_l has equal probability to be sampled from E_H and H. 5 values of $N \in \{1, 2, 3, 5, 10\}$ are considered. Fig. 6 with logarithmic time index shows the achieved average data rate when the sensing-probing-transmitting control is learned by Algorithm 1.

In Fig. 6, it can be observed that for different N values, Algorithm 1 converges to the same limit. As expected, larger N requires fewer learning steps to converge, which suggests a trade-off between computational load and learning speed. We also notice that there are turning points in all learning curves, explained as follows. An optimal algorithm expects to maximize the sum of immediate reward at the current slot and the expected reward in the future. Recall that, in Algorithm 1, we initially set $g_0(k) = 0, \forall k$. This means that at the beginning learning steps of Algorithm 1, the expected future reward is deemed zero. Thus, Algorithm 1 advises to maximize the immediate reward, i.e., act greedily for transmitting as many packets as possible. After a number of learning steps, expected future reward starts to take effect in Algorithm 1, and thus, the algorithm stops greedily sending packets and starts to jointly learn the sensing-probing-transmission policy. This causes temporal performance loss (as the algorithm may explore the system rather than transmitting as many packets as possible), but leads to the optimal performance asymptotically.

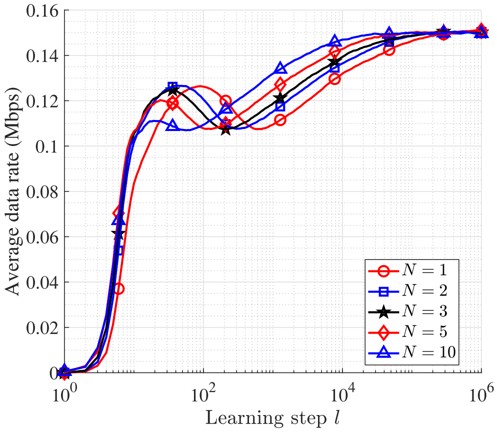


Fig. 6. Learning curves of Algorithm 1 under various N values.

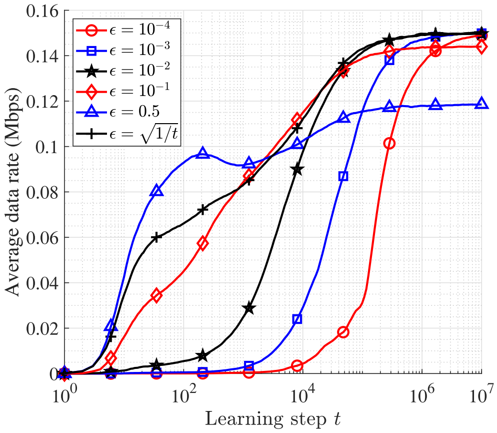


Fig. 7. Learning curves of Algorithm 2 under various exploration rates.

2) Online Learning Under Various Exploration Rate ϵ: With $N = 1$, we investigate the learning characteristics of Algorithm 2 for exploration rate $\epsilon \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 0.5\}$ and for ϵ adapted to t as $\epsilon = \sqrt{1/t}$. The average data rate is shown in Fig. 7.

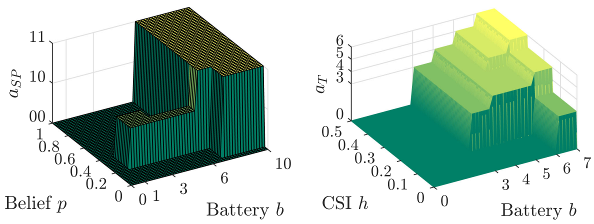
From Fig. 7, we see that larger ϵ tends to speed up learning. It is because larger ϵ implies more updates of CSI. However, too large ϵ can cause performance loss due to aggressive exploration. On the other hand, $\epsilon = \sqrt{1/t}$, which starts with large value and decreases over time, provides fast start-up and also almost-lossless asymptotic performance.

C. Structure of Learned Policy

After a learning algorithm converges, we can obtain a sensing-probing-transmitting policy from the learned function $g_L(\cdot)$. Specifically, given g_L, a policy is fully specified via a pair of sensing-probing 'sub-policy' $\hat{\pi}([b, p, e_H]|g_L)$ and transmitting 'sub-policy' $\hat{\pi}([b, h]|g_L)$ (recalling that $(\hat{\pi}(\cdot|g_L)$ is defined in (18)).

For $\mu_E = 5$, the learned sub-policies are shown in Fig. 8(a) and Fig. 8(b). Noting that the sensing-probing sub-policy is also

[8] There exists a tradeoff in setting τ_S and τ_P. With larger τ_S and τ_P, more accurate spectrum sensing and more accurate channel gain estimation can be achieved, at the cost of less time for transmission. As this paper focuses on design and analysis of a learning algorithm for the decision making process of the SU, we set $\tau_S = \tau_P = 1$ ms. Nevertheless, optimal selection of τ_S and τ_P is an interesting research topic, which can be investigated in future research work.



(a) Sensing-probing sub-policy. (b) Transmitting sub-policy.

Fig. 8. Learned sensing-probing and transmitting sub-policies.

a function of e_H, Fig. 8 shows the sensing-probing sub-policy with $e_H = 0$, for presentation simplicity. When $b > 7$, or when b is between 6 and 7 and belief p is more than a threshold, the optimal sensing-probing sub-policy is ‘11’ (i.e., to sense and to probe whenever possible). The sensing-probing action ‘10’ (sense but not probe) is selected when belief p is close to 0.4 and the battery level is less than 6, explained below. When belief p is close to 0.4, the SU is unsure about the channel availability, and thus, it is optimal to sense the channel to gain the channel availability knowledge and guide future decisions. On the other hand, the SU will not decide to probe, because when the battery level is less than 6, if the SU decides to probe, the total energy left for transmitting is less than $6 - e_S - e_P = 6 - 1 - 2 = 3$ (in other words, the SU will not be able to transmit, since the minimal transmit power level is 3 in our setting). For the transmitting sub-policy, higher battery level and/or higher channel gain h result in higher transmit power level, which is intuitive.

D. Performance Comparison

We next investigate the performance of learned policy. As the one-stage MDPs of [17], [18] are the most relevant works (see Section I), we compare our learned policy with a policy that is derived from a one-stage MDP. Since the works in [17], [18] assume static CSI, we implement the one-stage MDP of [17], [18] over a static channel with fixed channel gain being the average channel gain in our system.

We consider two baseline policies, namely “G-SPT” and “G-SP”. G-SPT is a purely greedy policy. Whenever the energy is sufficient, it senses and probes channel, and transmits at maximum power level. G-SP takes greedy action at sensing-probing stage, but adapts the transmit power based on probed CSI (i.e., G-SP is actually our proposed method in which there exists only one action ‘11’ for sensing-probing.)

We compare how these policies perform under different values of μ_E. First, we consider the ability of a policy to exploit channel access opportunities. This is measured by channel access probability, which is the probability that the channel is free while a sensing action is chosen. As a benchmark, this probability is upper bounded by the channel's idle probability $p_{01}/(p_{01}+p_{10}) = 0.1/0.2 = 0.5$. Fig. 9 shows the measured channel access probabilities of different policies, in which “1-stage” means one-stage MDP used in [17], [18]. Fig. 10 shows the data rate achieved by different policies, which is upper-bounded by $\tau_T/(\tau_S+\tau_P+\tau_T) \cdot p_{01}/(p_{01}+p_{10}) \cdot p_O \cdot \mathbb{E}[W \log_2(1+e_T^M h_f)] \approx 0.78$ Mbps, where $e_T^M = \max\{\mathbf{E_T}\} = 6$.

It can be seen that, when μ_E increases, all the policies have more channel access opportunities and higher data rates. And with a high enough energy supply, channel access probabilities of all the policies achieve the upper bound.

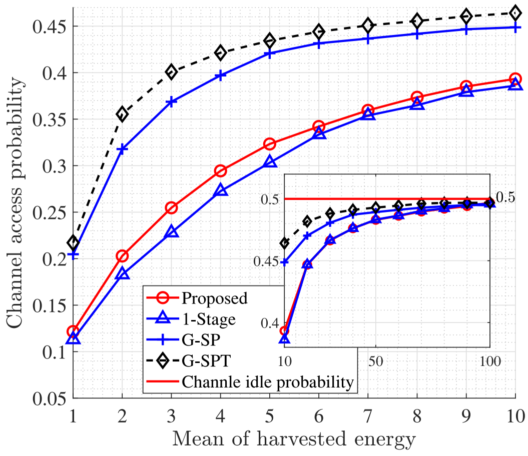


Fig. 9. Channel access probability under different μ_E.

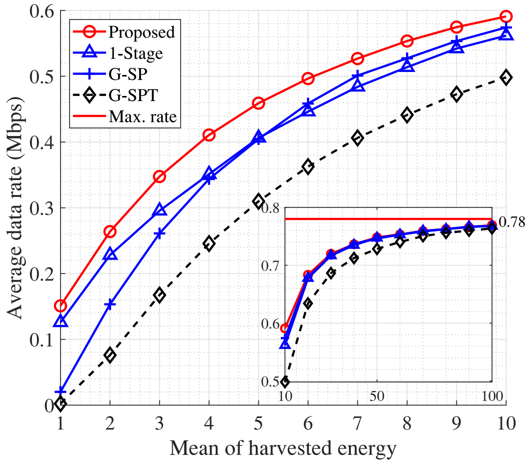


Fig. 10. Data rates for different μ_E.

Note that G-SPT and G-SP are efficient at exploiting channel access opportunities because they are aggressive in sensing and probing. However, their greedy actions may result in a lack of energy for data transmission, especially when harvested energy is limited. Therefore, when μ_E is small, their achieved data rates are small. Nevertheless, G-SP's data rate is higher than that of G-SPT. The main reason is that G-SP adapts transmit power based on channel fading status, and therefore, uses energy more efficiently for data transmitting than a purely greedy policy.

As for one-stage MDP, it achieves higher data rate than G-SP, when $\mu_E < 5$; but is slightly inferior to G-SP when $\mu_E > 5$. That is because, when energy is limited, making proper sensing and probing decision to save energy for transmission is more important. However, given a large μ_E, energy expenses due to greedy sensing and probing action are marginal relative to the available energy, while properly selecting transmission energy is of more importance. Therefore, with a large μ_E, G-SP, which makes transmission decision based on instantaneous CSI, demonstrates better performance than one-stage MDP.

With full adaptation and a two-stage decision scheme, our proposed method achieves favorable energy tradeoff between sensing-probing and transmission stages, and thus, achieves the best data rate.

VII. Conclusions

This paper studied the optimal sensing, probing and power control problem for an energy harvesting SU operating in a fading channel. The problem was modeled as a two-stage continuous state MDP and then simplified via the after-state value function. The SU learns this function without knowledge of statistical distributions of the wireless channel and the energy harvesting process. For this learning process, we developed a reinforcement learning algorithm and investigated its learning characteristics and performance via simulation.

Our work can be extended to the following scenarios.

Multiple-channel scenario: With multiple channels, the SU can maintain a belief value for availability of each channel. When the SU decides to sense, it needs to decide which channel to sense. Similar to our work, a two-stage MDP can be developed, and after-state formulation can be used to find the optimal policy.

Multiple-SU scenario: With multiple SUs, if an SU decides to probe the channel, it applies a contention procedure such as carrier sense multiple access. If it wins the contention, it probes the channel; otherwise, it keeps idle until the next time slot. Accordingly, by adding a probability of successful contention into our two-stage MDP, our algorithms can be applied to get optimal policy for each SU.

Bursty-traffic scenario: Now the data buffer fluctuates randomly. Therefore, not only the amount of transmitted data, but also the reduction of packet losses due to data buffer overflow is of interest. Thus, we can include the occupancy of data buffer into our definition of "state", and redefine the reward function as a weighted combination of sent data with a positive weight and data buffer occupancy with a negative weight (such a reward definition is also considered in [34]). We can then formulate a two-stage MDP, and use after-state value function to find the optimal policy.

Appendix A
Proof of Theorem 1

Theorem 1 is proved with the use of contraction theory. Specifically, we show the solution to (14) uniquely exists, and it is the fixed point of a contraction mapping. Furthermore, the value iteration algorithm (15) converges to the fixed point.

First, define the set of bounded functions $J : \mathbb{D} \to \mathbb{R}$ as $\mathbb{F}$. Then, let T^* be an operator on $\mathbb{F}$, and for any $J \in \mathbb{F}$, T^*J is another function with domain $\mathbb{D}$, whose value at β is defined as

$$(T^*J)(\beta) = \gamma \mathop{\mathbb{E}}_{X'|\beta}\left[\max_{a' \in \mathbb{A}([\beta, X'])}\left\{r([\beta, X'], a') + \sum_{i=1}^{\mathcal{N}(a')} p_i(\beta, a')J(\varrho_i([\beta, X'], a'))\right\}\right].$$

By Assumption 1, it is easy to check that, given J is bounded, T^*J is bounded (i.e., $T^*J \in \mathbb{F}$). Therefore, T^* is a mapping from $\mathbb{F}$ to $\mathbb{F}$. It is shown in [35, p. 211] that $\mathbb{F}$ is complete under the maximum norm. Furthermore, as shown in the following, T^* is a contraction mapping under the maximum norm with modulus γ. Therefore, the contraction theory applies to T^*.

Due to the contraction theory [35, p. 209], there exists a unique fixed point for T^*, denoted as J^*, such that $T^*J^* = J^*$, i.e., function J^* does not change under operator T^*. Note that equation $T^*J^* = J^*$ is exactly the after-state Bellman equation (14). Therefore, we have shown that there is a unique solution to (14).

In addition, the contraction theory [35, p. 209] states that, for arbitrary function $J_0 \in \mathbb{F}$, $\lim_{l\to\infty} T^{*l}J_0 = J^*$. Note that $T^{*l}J_0$ means the function that is generated by, starting from J_0, iteratively applying operator T^* on previously generated function for l times, which exactly describes the value iteration algorithm (15). This has proved the value iteration algorithm (15) converges to J^*.

Hence, there only remains to show that T^* is a contraction mapping. Given any two functions $J_1, J_2 \in \mathbb{F}$, for β that satisfies $(T^*J_1)(\beta) \geq (T^*J_2)(\beta)$, we have

$$\begin{aligned}
0 &\leq (T^*J_1)(\beta) - (T^*J_2)(\beta) \\
&= \gamma \mathop{\mathbb{E}}_{X'|\beta}\left[\max_{a_1 \in \mathbb{A}([\beta, X'])}\left\{r([\beta, X'], a_1) + \sum_{i=1}^{\mathcal{N}(a_1)} p_i(\beta, a_1)J_1(\rho_i([\beta, X'], a_1))\right\}\right. \\
&\qquad \left. - \max_{a_2 \in \mathbb{A}([\beta, X'])}\left\{r([\beta, X'], a_2) + \sum_{i=1}^{\mathcal{N}(a_2)} p_i(\beta, a_2)J_2(\rho_i([\beta, X'], a_2))\right\}\right] \\
&\leq \gamma \mathop{\mathbb{E}}_{X'|\beta}\left[r([\beta, X'], a_1^*) + \sum_{i=1}^{\mathcal{N}(a_1^*)} p_i(\beta, a_1^*)J_1(\rho_i([\beta, X'], a_1^*))\right. \\
&\qquad \left. - r([d, X'], a_1^*) - \sum_{i=1}^{\mathcal{N}(a_1^*)} p_i(\beta, a_1^*)J_2(\rho_i([\beta, X'], a_1^*))\right] \\
&= \gamma \mathop{\mathbb{E}}_{X'|\beta}\left[\sum_{i=1}^{\mathcal{N}(a_1^*)} p_i(\beta, a_1^*) \times \left(J_1(\rho_i([\beta, X'], a_1^*)) - J_2(\rho_i([d, X'], a_1^*))\right)\right] \\
&\leq \gamma \mathop{\mathbb{E}}_{X'|\beta}\left[\sum_{i=1}^{\mathcal{N}(a_1^*)} p_i(a_1^*)||J_1 - J_2||\right] = \gamma||J_1 - J_2||, \qquad (27)
\end{aligned}$$

where

$$a_1^* = \mathop{\arg\max}_{a_1 \in \mathbb{A}([\beta, X'])}\left\{r([\beta, X'], a_1) + \sum_{i=1}^{\mathcal{N}(a_1)} p_i(\beta, a_1)J_1(\rho_i([\beta, X'], a_1))\right\},$$

and $||\cdot||$ is the maximum norm.

For β that satisfies $(T^*J_1)(\beta) < (T^*J_2)(\beta)$, we can get

$$0 < (T^*J_2)(\beta) - (T^*J_1)(\beta) \leq \gamma||J_1 - J_2||, \qquad (28)$$

following similar procedure by replacing J_1 to J_2, and vice versa. Therefore, combining (27) with (28) gives $|(T^*J_1)(\beta) - (T^*J_2)(\beta)| \leq \gamma||J_1 - J_2||$ for all $\beta \in \mathbb{D}$, i.e., $||T^*J_1 - T^*J_2|| \leq \gamma||J_1 - J_2||$. It has proved that T^* is a contraction mapping on $\mathbb{F}$ with modulus γ. And the proof of Theorem 1 is completed.

Appendix B
Proof of Theorem 2

With $s = [d, x]$, define a function

$$G(s) \triangleq \max_{a \in \mathbb{A}(s)} \left\{ r(s,a) + \sum_{i=1}^{\mathcal{N}(a)} p_i(d,a) J^*(\varrho_i(s,a)) \right\}. \tag{29}$$

Expanding $J^*(\varrho(s,a))$ from equation (14) gives

$$\begin{aligned} G(s) = & \max_{a \in \mathbb{A}([d,x])} \left\{ r(s,a) + \sum_{i=1}^{\mathcal{N}(a)} p_i(d,a) \right. \\ & \times \gamma \mathop{\mathbb{E}}_{X'|\varrho_i(s,a)} \left[\max_{a' \in \mathbb{A}([\varrho_i(s,a), X'])} \left\{ r(\varrho_i(s,a), X', a') \right. \right. \\ & \left. \left. \left. + \sum_{j=1}^{\mathcal{N}(a')} p_j(d', a') J(\varrho_j(\varrho_i(s,a), X', a')) \right\} \right] \right\} \\ = & \max_{a \in \mathbb{A}([d,x])} \left\{ r(s,a) \right. \\ & \left. + \gamma \sum_{i=1}^{\mathcal{N}(a)} p_i(d,a) \mathop{\mathbb{E}}_{X'|\varrho_i(s,a)} [G(\varrho_i(s,a), X')] \right\} \\ = & \max_{a \in \mathbb{A}([d,x])} \left\{ r(s,a) + \gamma \mathbb{E}\left[G(S') | s, a \right] \right\}, \end{aligned} \tag{30}$$

where the definition of G implies the second equality, and (12) implies the last equality. Note that (30) is exactly the state Bellman equation (9). Therefore, function $G = V^*$ solves (9), and the relationship (16) is established. Finally, with (29) and the definition of the after-state Bellman equation (14), the relationship (17) is established, which completes the proof.

Appendix C
Proof of Theorem 3

For Algorithm 1, we define two operators H and $\hat{H}$. Let H be an operator on functions $\mathbb{K} \mapsto \mathbb{R}$. Applying H on a function g, i.e., Hg, gives another function with domain $\mathbb{K}$, and its value at k is defined as

$$\begin{aligned} (Hg)(k) = \gamma \mathop{\mathbb{E}}_{X'|q(k)} & \left[\max_{a' \in \mathbb{A}([q(k), X'])} \left\{ r([q(k), X'], a') \right. \right. \\ & \left. \left. + \sum_{i=1}^{\mathcal{N}(a')} p_i(q(k), a') g(\omega(\varrho_i([q(k), X'], a'))) \right\} \right]. \end{aligned}$$

Similarly, define another operator on functions $\mathbb{K} \mapsto \mathbb{R}$ as

$$\begin{aligned} (\hat{H}g)(k) = \gamma & \max_{a' \in \mathbb{A}([q(k), X'])} \left\{ r([q(k), X'], a') \right. \\ & \left. + \sum_{i=1}^{\mathcal{N}(a')} p_i(q(k), a') g(\omega(\varrho_i([q(k), X'], a'))) \right\}, \end{aligned}$$

where X' is a r.v. with pdf $f_X(\cdot|q(k))$. Note that the outcome of $\hat{H}g$ is random, and depends on the realization of X'.

Note that, in Algorithm 1, at any iteration l, $g_l(k)$ does not change for $k \notin \bar{K}_l$. Therefore, the step size value $\alpha_l(k)$, $\forall k \notin \bar{K}_l$, does not affect the algorithm. By defining $\alpha_l(k) = 0$, $\forall k \notin \bar{K}_l$, and with the operators H and $\hat{H}$, the updating (19) can be rewritten as, $\forall k \in \mathbb{K}$:

$$g_{l+1}(k) = (1 - \alpha_l(k)) g_l(k) + \alpha_l(k)((Hg_l)(k) + w_l(k)) \tag{31}$$

where $w_l(k) = (\hat{H}g_l)(k) - (Hg_l)(k)$.

A. *Proof of Statement (i)*

From [36, Proposition 4.4], we have following lemma.

Lemma 1: Given following conditions,

a) H is a contraction mapping under maximum norm;
b) for all k, $\sum_{l=0}^{\infty} \alpha_l(k) = \infty$, and $\sum_{l=0}^{\infty} \alpha_l^2(k) < \infty$;
c) for all k and l, $\mathbb{E}[w_l(k)|g_l] = 0$;
d) there exist constant C_1 and C_2 such that $\mathbb{E}[w_l^2(k)|g_l] \leq C_1 + C_2||g_l||^2$;

the sequence of functions $\{g_l\}_l$ generated from iteration (31) converges to a function g_∞ with probability 1, and the limiting function g_∞ satisfying $Hg_\infty = g_\infty$.

We prove the statement (i) of Theorem 3 by checking the four conditions of Lemma 1 as follows. First, the contraction mapping condition (a) of H can be established in a similar procedure as the proof of Theorem 1, and is omitted here. Then, due to assumptions (24) and (25) of Theorem 3, the condition (b) about α_l is satisfied. In addition, we have $\mathbb{E}[w_l(k)|g_l] = 0$ via the definition of H and $\hat{H}$. Therefore, the condition (c) is satisfied. Finally, we have to prove the condition (d): the bounded variance property of w_l. For given k and l, we define a function as

$$\begin{aligned} I(x) = \gamma & \max_{a' \in \mathbb{A}([q(k), x])} \left\{ r([q(k), x], a') \right. \\ & \left. + \sum_{i=1}^{\mathcal{N}(a')} p_i(q(k), a') g_l(\omega(\varrho_i([q(k), x], a'))) \right\}. \end{aligned}$$

With the notation $I(x)$, we have

$$\begin{aligned} \mathbb{E}[w_l^2(k)|g_l] & = \mathop{\mathbb{E}}_{X'|q(k)} \left[\left(I(X') - \mathop{\mathbb{E}}_{Y'|q(k)} [I(Y')] \right)^2 \middle| g_l \right] \\ & = \mathop{\mathbb{E}}_{X'|q(k)} \left[\left(\mathop{\mathbb{E}}_{Y'|q(k)} [I(X') - I(Y')] \right)^2 \middle| g_l \right] \\ & \leq \mathop{\mathbb{E}}_{X'|q(k)} \left[\left(\mathop{\mathbb{E}}_{Y'|q(k)} [2 \max\{ |I(X')|, |I(Y')| \}] \right)^2 \middle| g_l \right] \end{aligned}$$

$$\begin{aligned}
&\leq \mathop{\mathbb{E}}_{X'|q(k)}\left[\left(\mathop{\mathbb{E}}_{Y'|q(k)}[2|I(X')|]\right)^2\Big|g_l\right] \\
&\quad + \mathop{\mathbb{E}}_{X'|q(k)}\left[\left(\mathop{\mathbb{E}}_{Y'|q(k)}[2|I(Y')|]\right)^2\Big|g_l\right] \\
&\overset{\textcircled{1}}{\leq} \mathop{\mathbb{E}}_{X'|q(k)}\left[(2|I(X')|)^2|g_l\right] + \mathop{\mathbb{E}}_{X'|q(k)}\left[(2L_1+2||g_l||)^2|g_l\right] \\
&\overset{\textcircled{2}}{\leq} 8L_2+8||g_l||^2+8L_1^2+8||g_l||^2 = 8(L_2+L_1^2)+16||g_l||^2,
\end{aligned}$$

where the inequalities ①and ②hold from Assumption 1 and the fact that $(x+y)^2 \leq 2x^2+2y^2$ for any real value x and y. Therefore, it is proven that $\mathbb{E}[w_l^2(k)|g_l]$ is bounded by $8(L_2+L_1^2)+16||g_l||^2$, which completes the proof of the statement (i) of Theorem 3.

B. *Proof of Statement (ii)*

First, define a partial order for functions $\mathbb{K} \mapsto \mathbb{R}$ as follows. If $g_1(k) \leq g_2(k)$, $\forall k$, we say $g_1 \leq g_2$. It is easy to check that, given any two functions g_1 and g_2 satisfying $g_1 \leq g_2$, we have $Hg_1 \leq Hg_2$.

Then, define a function $\bar{g}(k) \triangleq \inf_{\beta \in \mathbb{D}(k)} J^*(\beta) + \frac{\xi}{1-\gamma}$. Applying H on $\bar{g}$ gives

$$\begin{aligned}
(H\bar{g})(k) &= \gamma \mathop{\mathbb{E}}_{X'|q(k)}\Bigg[\max_{a' \in \mathbb{A}([q(k),X'])}\Big\{r([q(k),X'],a') \\
&\quad + \sum_{i=1}^{\mathcal{N}(a')} p_i(q(k),a')\bar{g}(\omega(\varrho_i([q(k),X'],a')))\Big\}\Bigg] \\
&\overset{\textcircled{3}}{\leq} \gamma \mathop{\mathbb{E}}_{X'|q(k)}\Bigg[\max_{a' \in \mathbb{A}([q(k),X'])}\Bigg\{r([q(k),X'],a') \\
&\quad + \sum_{i=1}^{\mathcal{N}(a')} p_i(q(k),a')\left(J^*(\varrho_i([q(k),X'],a')) + \frac{\xi}{1-\gamma}\right)\Bigg\}\Bigg] \\
&\overset{\textcircled{4}}{=} J^*(q(k)) + \frac{\gamma\xi}{1-\gamma} \overset{\textcircled{5}}{\leq} \inf_{\beta \in \mathbb{D}(k)} J^*(\beta) + \xi + \frac{\gamma\xi}{1-\gamma} = \bar{g}(k),
\end{aligned}$$

where inequality ③ is due to the definition of $\bar{g}(k)$, equality ④ comes from the after-state Bellman equation (14), and inequality ⑤ is due to the definition of ξ in (22). Therefore, we have $(H\bar{g})(k) \leq \bar{g}(k)$ for all k, i.e., $H\bar{g} \leq \bar{g}$.

Combining the fact that $Hg_1 \leq Hg_2$, if $g_1 \leq g_2$, with the fact that $H\bar{g} \leq \bar{g}$, we have $H^k\bar{g} \leq \bar{g}$, where H^k means applying H operator k times. Then, due to Lemma 1 in the proof of statement (i), we have $\lim_{k\to\infty} H^k\bar{g} = g_\infty \leq \bar{g}$, which means $g_\infty(k) \leq \inf_{\beta\in\mathbb{D}(k)} J^*(\beta) + \frac{\xi}{1-\gamma}, \forall\, k$. Therefore, we get $J^*(\beta) \geq g_\infty(\omega(\beta)) - \frac{\xi}{1-\gamma}$, $\forall\, \beta$. From the definition of J_∞ in (26), $J^*(\beta) - J_\infty(\beta) \geq -\frac{\xi}{1-\gamma}, \forall\, \beta$, follows.

On the other hand, defining $\underline{g}(k) = \sup_{\beta\in\mathbb{D}(k)} J^*(\beta) - \frac{\xi}{1-\gamma}$ and following the similar procedure, we can prove $H\underline{g} \geq \underline{g}$, and therefore, get $J^*(\beta) \leq g_\infty(\omega(\beta)) + \frac{\xi}{1-\gamma}$. In turn, it implies $J^*(\beta) - J_\infty(\beta) \leq \frac{\xi}{1-\gamma}$, which completes the proof of statement (ii) in Theorem 3.

C. *Proof of Statement (iii)*

For any policy π, define an operator T^π on $\mathbb{F}$ ($\mathbb{F}$ is defined in Appendix A) as

$$\begin{aligned}
(T^\pi J)(\beta) = \gamma \mathop{\mathbb{E}}_{X'|\beta}\Bigg[&r([\beta,X'],\pi) \\
&+ \sum_{i=1}^{\mathcal{N}(\pi)} p_i(\beta,\pi)J(\varrho_i([\beta,X'],\pi))\Bigg],
\end{aligned} \tag{32}$$

with π inside r, $\mathcal{N}$, p_i and ϱ_i denoting $\pi([\beta,X'])$. And from the state transition kernel (12), J^{π_∞} as defined by (23) can be recursively rewritten as $J^{\pi_\infty}(\beta) = \gamma\,\mathbb{E}_{X'|\beta}[r([\beta,X'],\pi_\infty) + \sum_{i=1}^{\mathcal{N}(\pi_\infty)} p_i(\beta,\pi_\infty)J^{\pi_\infty}(\varrho_i([\beta,X'],\pi_\infty))]$. By comparing this expression with T^π in (32), we have

$$T^{\pi_\infty}J^{\pi_\infty} = J^{\pi_\infty}. \tag{33}$$

In addition, similar to the proof of Theorem 1, T^π is a contraction mapping with modulus γ, which means

$$||T^{\pi_\infty}J_1 - T^{\pi_\infty}J_2|| \leq \gamma||J_1 - J_2|| \tag{34}$$

for any J_1 and J_2. Besides, from the definitions of $\hat{\pi}(\cdot|g_\infty)$ (i.e., π_∞) in (18) and J_∞ in (26), we have (T^* defined in Appendix A)

$$T^{\pi_\infty}J_\infty = T^*J_\infty. \tag{35}$$

Furthermore, from statement (ii) of Theorem 3, we have

$$||J^* - J_\infty|| \leq \frac{\xi}{1-\gamma}. \tag{36}$$

Finally, it is shown in the proof of Theorem 1 that

$$T^*J^* = J^*, \tag{37}$$

$$||T^*J_1 - T^*J_2|| \leq \gamma||J_1 - J_2||, \text{for any } J_1 \text{ and } J_2. \tag{38}$$

By combining the above results, we have

$$\begin{aligned}
||J^{\pi_\infty} - J^*|| &\overset{\textcircled{6}}{=} ||T^{\pi_\infty}J^{\pi_\infty} - J^*|| \\
&\overset{\textcircled{7}}{\leq} ||T^{\pi_\infty}J^{\pi_\infty} - T^{\pi_\infty}J_\infty|| + ||T^{\pi_\infty}J_\infty - J^*|| \\
&\overset{\textcircled{8}}{\leq} \gamma||J^{\pi_\infty} - J_\infty|| + ||T^*J_\infty - T^*J^*|| \\
&\overset{\textcircled{9}}{\leq} \gamma||J^{\pi_\infty} - J^*|| + \gamma||J^* - J_\infty|| + \gamma||J_\infty - J^*|| \\
&\overset{\textcircled{A}}{\leq} \gamma||J^{\pi_\infty} - J^*|| + \frac{2\gamma\xi}{1-\gamma},
\end{aligned} \tag{39}$$

where ⑥ is from (33); ⑦ is the triangle inequality; ⑧ is from (34), (35) and (37); ⑨ is from the triangle inequality and (38), and Ⓐ is from (36). Finally, from (39), we have $||J^{\pi_\infty} - J^*|| \leq \frac{2\gamma\xi}{(1-\gamma)^2}$, which proves statement (iii) of Theorem 3.

References

[1] K. Wu, H. Jiang, and C. Tellambura, "Sensing, probing, and transmitting strategy for energy harvesting cognitive radio," in *Proc. IEEE Int. Conf. Commun.*, May 2017, pp. 1–6.

[2] M. L. Ku, W. Li, Y. Chen, and K. J. R. Liu, "Advances in energy harvesting communications: Past, present, and future challenges," *IEEE Commun. Surv. Tut.*, vol. 18, no. 2, pp. 1384–1412, Apr.–Jun. 2016.

[3] S. Haykin, "Cognitive radio: Brain-empowered wireless communications," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 2, pp. 201–220, Feb. 2005.

[4] S. Park, H. Kim, and D. Hong, "Cognitive radio networks with energy harvesting," *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, pp. 1386–1397, Mar. 2013.

[5] T. Shu and M. Krunz, "Throughput-efficient sequential channel sensing and probing in cognitive radio networks under sensing errors," in *Proc. 15th Annu. Int. Conf. Mobile Comput. Netw.*, Sep. 2009, pp. 37–48.

[6] S. S. Tan, J. Zeidler, and B. Rao, "Opportunistic channel-aware spectrum access for cognitive radio networks with interleaved transmission and sensing," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2376–2388, May 2013.

[7] T. Cui and C. Tellambura, "Joint data detection and channel estimation for OFDM systems," *IEEE Trans. Commun.*, vol. 54, no. 4, pp. 670–679, Apr. 2006.

[8] G. Wang, F. Gao, Y.-C. Wu, and C. Tellambura, "Joint CFO and channel estimation for OFDM-based two-way relay networks," *IEEE Trans. Wireless Commun.*, vol. 10, no. 2, pp. 456–465, Feb. 2011.

[9] G. Wang, F. Gao, W. Chen, and C. Tellambura, "Channel estimation and training design for two-way relay networks in time-selective fading environments," *IEEE Trans. Wireless Commun.*, vol. 10, no. 8, pp. 2681–2691, Aug. 2011.

[10] S. Park and D. Hong, "Optimal spectrum access for energy harvesting cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 12, pp. 6166–6179, Dec. 2013.

[11] W. Chung, S. Park, S. Lim, and D. Hong, "Spectrum sensing optimization for energy-harvesting cognitive radio systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 5, pp. 2601–2613, May 2014.

[12] D. T. Hoang, D. Niyato, P. Wang, and D. I. Kim, "Performance optimization for cooperative multiuser cognitive radio networks with RF energy harvesting capability," *IEEE Trans. Wireless Commun.*, vol. 14, no. 7, pp. 3614–3629, Jul. 2015.

[13] Pratibha, K. H. Li and K. C. Teh, "Dynamic cooperative sensing-access policy for energy-harvesting cognitive radio systems," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10137–10141, Dec. 2016.

[14] A. Celik, A. Alsharoa, and A. E. Kamal, "Hybrid energy harvesting-based cooperative spectrum sensing and access in heterogeneous cognitive radio networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 3, no. 1, pp. 37–48, Mar. 2017.

[15] D. Zhang, Z. Chen, M. K. Awad, N. Zhang, H. Zhou, and X. S. Shen, "Utility-optimal resource management and allocation algorithm for energy harvesting cognitive radio sensor networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3552–3565, Dec. 2016.

[16] C. Xu, M. Zheng, W. Liang, H. Yu, and Y.-C. Liang, "End-to-end throughput maximization for underlay multi-hop cognitive radio networks with RF energy harvesting," *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3561–3572, Jun. 2017.

[17] A. Sultan, "Sensing and transmit energy optimization for an energy harvesting cognitive radio," *IEEE Wireless Commun. Lett.*, vol. 1, no. 5, pp. 500–503, Oct. 2012.

[18] Z. Li, B. Liu, J. Si, and F. Zhou, "Optimal spectrum sensing interval in energy-harvesting cognitive radio networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 3, no. 2, pp. 190–200, Jun. 2017.

[19] S. Yin, Z. Qu, and S. Li, "Achievable throughput optimization in energy harvesting cognitive radio systems," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 3, pp. 407–422, Mar. 2015.

[20] Pratibha, K. H. Li and K. C. Teh, "Optimal spectrum access and energy supply for cognitive radio systems with opportunistic RF energy harvesting," *IEEE Trans. Veh. Technol.*, vol. 66, no. 8, pp. 7114–7122, Aug. 2017.

[21] D. Zhang *et al.*, "Energy-harvesting-aided spectrum sensing and data transmission in heterogeneous cognitive radio sensor network," *IEEE Trans. Veh. Technol.*, vol. 66, no. 1, pp. 831–843, Jan. 2017.

[22] J. J. Pradha, S. S. Kalamkar, and A. Banerjee, "Energy harvesting cognitive radio with channel-aware sensing strategy," *IEEE Commun. Lett.*, vol. 18, no. 7, pp. 1171–1174, Jul. 2014.

[23] A. Goldsmith, S. A. Jafar, I. Maric, and S. Srinivasa, "Breaking spectrum gridlock with cognitive radios: An information theoretic perspective," *Proc. IEEE*, vol. 97, no. 5, pp. 894–914, May 2009.

[24] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York, NY, USA: Wiley, 1994.

[25] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998.

[26] S. Geirhofer, L. Tong, and B. M. Sadler, "A measurement-based model for dynamic spectrum access in WLAN channels," in *Proc. IEEE Military Commun. Conf.*, Oct. 2006, pp. 1–7.

[27] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 3, pp. 589–600, Apr. 2007.

[28] S. Ulukus *et al.*, "Energy harvesting wireless communications: A review of recent advances," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 3, pp. 360–381, Mar. 2015.

[29] S. Luo, R. Zhang, and T. J. Lim, "Optimal save-then-transmit protocol for energy harvesting wireless transmitters," *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, pp. 1196–1207, Mar. 2013.

[30] H. Li, C. Huang, P. Zhang, S. Cui, and J. Zhang, "Distributed opportunistic scheduling for energy harvesting based wireless networks: A two-stage probing approach," *IEEE/ACM Trans. Netw.*, vol. 24, no. 3, pp. 1618–1631, Jun. 2016.

[31] C. R. Stevenson, G. Chouinard, Z. Lei, W. Hu, S. J. Shellhammer, and W. Caldwell, "IEEE 802.22: The first cognitive radio wireless regional area network standard," *IEEE Commun. Mag.*, vol. 47, no. 1, pp. 130–138, Jan. 2009.

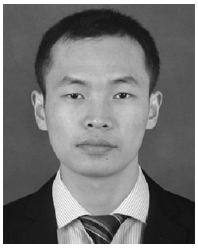
[32] H. Kushner and G. G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*. New York, NY, USA: Springer, 2003.

[33] J. Seguro and T. Lambert, "Modern estimation of the parameters of the Weibull wind speed distribution for wind energy analysis," *J. Wind Eng. Ind. Aerodynamics*, vol. 85, no. 1, pp. 75–84, Mar. 2000.

[34] K. Prabuchandran, S. K. Meena, and S. Bhatnagar, "Q-learning based energy management policies for a single sensor node with finite buffer," *IEEE Commun. Lett.*, vol. 2, no. 1, pp. 82–85, Feb. 2013.

[35] D. P. Bertsekas, *Abstract Dynamic Programming*. Belmont, MA, USA: Athena Scientific, 2013.

[36] D. P. Bertsekas and J. N. Tsitsiklis, *Neuro-Dynamic Programming*. Belmont, MA, USA: Athena Scientific, 1996.



Keyu Wu received the Ph.D. degree from the University of Alberta, Canada, in 2018. His research interests include detection and estimation, dynamic stochastic control, machine learning, and their application in wireless communications.

Hai Jiang (SM'15) received the Ph.D. degree in electrical engineering from the University of Waterloo, Waterloo, Ontario, Canada, in 2006. Since July 2007, he has been a Faculty Member with the University of Alberta, Edmonton, AB, Canada, where he is currently a Professor with the Department of Electrical and Computer Engineering. His research interests include radio resource management, cognitive radio networking, and cooperative communications.

Chintha Tellambura (F'11) received the Ph.D. degree in electrical engineering from the University of Victoria, Victoria, BC, Canada.

He was with the Monash University, Australia, from 1997 to 2002. He is currently a Professor with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada. His current research interests include the designs, modeling, and analysis of cognitive radios, heterogeneous cellular networks, 5G wireless networks, and machine learning algorithms.