Optimal Online Transmission Policy for Energy-Constrained Wireless-Powered Communication Networks

Xian Li[†], Xiangyun Zhou[‡], Derrick Wing Kwan Ng[§], and Changyin Sun[†]

[†]School of Automation, Southeast University, Nanjing, China
[‡] Research School of Engineering, The Australian National University, Canberra, ACT, Australia
[§]School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, NSW, Australia
Email: seulixian@gmail.com, xiangyun.zhou@anu.edu.au, w.k.ng@unsw.edu.au, cysun@seu.edu.cn

***Abstract*—This work considers the design of online transmission policy in a wireless-powered communication system with a given energy budget. The system design objective is to maximize the long-term throughput of the system exploiting the energy storage capability at the wireless-powered node. We formulate the design problem as a constrained Markov decision process (CMDP) problem and obtain the optimal policy of transmit power and time allocation in each fading block via the Lagrangian approach. To investigate the system performance in different scenarios, numerical simulations are conducted with various system parameters. Our simulation results show that the optimal policy significantly outperforms a myopic policy which only maximizes the throughput in the current fading block. Moreover, the optimal allocation of transmit power and time is shown to be insensitive to the change of modulation and coding schemes, which facilitates its practical implementation.**

I. Introduction

Wireless-powered communication networks (WPCNs), which usually consist of a hybrid access point (H-AP) and several user equipments (UEs) [1], have drawn significant attention recently. The system performance in terms of different metrics (e.g., throughput [2], outage [3], energy efficiency [4]) for various scenarios (e.g., point-to-point [5], two-hop relaying [6], multiple-input and multiple-output (MIMO) [7]) have been thoroughly investigated. However, most existing works devoted their efforts to studying the system performance of only one time block (slot), where all the harvested energy is exhausted immediately without exploiting long-term energy storage. In practice, due to the variability of the communication channel quality, it is more reasonable to store part of or even all the harvested energy in the battery when the channel undergoes deep fading. Thus it is of great importance to study the transmission policy for optimizing long-term system performance with long-term energy storage capability.

Some research efforts have been devoted to improving the long-term system performance. Considering two simple online transmission policies for a single-user WPCN, the limiting distribution of the stored energy at the UE as well as the outage performance of the system was investigated in [8]. In [9], the data rate maximization problem of an orthogonal frequency division multiplexing (OFDM)-based WPCN was studied. To jointly optimize the subchannel allocation and the power allocation over time, an offline algorithm and an online algorithm were designed for the case of non-causal channel state information (CSI) and causal CSI, respectively. Considering the variation of the CSI and the evolution of the battery state over slots, the long-term system performance of a two-user WPCN in an infinite horizon was studied in [10]. Based on the theory of Markov decision process, the optimal online policy was obtained to maximize the long-term system throughput. After that, the authors in [11] extended this work to a full-duplex scenario where the H-AP transfers energy and receives information data simultaneously. The corresponding optimal online policy for the full-duplex case was obtained and the long-term performance gap between the full-duplex WPCN and the half-duplex WPCN was also discussed. However, the temporal correlation of the time-varying channels, which can be exploited to improve the system performance, was not considered in these works. Also, the H-AP in these works, e.g., [8]–[11], was assumed to equip with an infinite power supply and hence energy consumption of the system has not been a consideration in the previous studies.

In this paper, we focus on the long-term throughput performance of a WPCN with limited system energy budget. More specifically, considering the H-AP with a finite amount of energy, we design an optimal online transmission policy to maximize the throughput over an infinite horizon. The contribution of the work lies in both the modeling and solution development of the throughput maximization problem. First, during problem formulation, the finite state Markov channel (FSMC) model is adopted to capture the temporal-correlation behavior of the fading channel. Moreover, practical aspects including circuit power consumption and efficiency of the power amplifier are considered to evaluate the total system energy consumption. Then, we formulate the problem as a constrained Markov decision process (CMDP) problem and solve it optimally via the Lagrangian approach, where a bisection search is introduced to update the corresponding Lagrange multiplier. Subsequently, the long-term system performance under various scenario is studied via numerical simulations. In particular, the impact of the system parameters on the system performance is thoroughly discussed, which provides practical insights on the design and implementation of the WPCN.

II. System model

As shown in Fig. 1, we consider a WPCN consisting of a H-AP and a single-antenna UE in this paper. The H-AP is equipped with a directional antenna and the UE is driven by a rechargeable battery with maximum capacity $B_{\max}$. A time-correlated block fading channel is considered between the H-AP and the UE, where the channel power gain remains constant in a block but varies from one to another. In block $t \in 1, 2, \cdots$, the channel power gain is expressed as $H_t = \theta_t d^{-\alpha}$, where θ_t is a random variable capturing the multipath fading, d is the distance between the H-AP and the UE, and α is the path loss exponent. In each block, there is a wireless energy transfer (WET) period and a wireless information transfer (WIT) period. The UE first harvests energy from the H-AP and stores it in the battery during WET, and then transmits its data to the H-AP utilizing the energy stored in the battery during the following WIT.

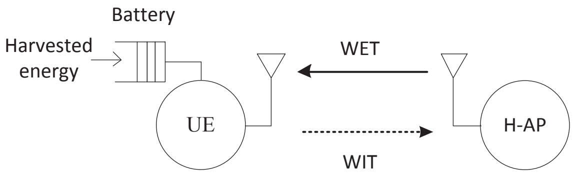

Fig. 1. The system model of a WPCN.

In this paper, we aim at maximizing the system throughput over an infinite horizon under a given energy budget constraint. This considered problem can be formulated in the framework of a CMDP which consists five elements: the system state space $\mathcal{S}$, the action space $\mathcal{A}$, the probability transition matrix $\mathcal{P}$, the reward function $r(\cdot)$, and the cost function $e(\cdot)$. In the following, detail descriptions of these five elements are provided.

A. System States

For the considered system, the optimal policy is constructed at the H-AP based on the channel information and the battery information. We assume that in the current block, perfect CSI as well as the UE's battery information is available at the H-AP (In practice, this information can be acquired in the training phase at the beginning of each block). Correspondingly, in block t, the system state $\boldsymbol{s}_t \in \mathcal{S}$ consists the channel state $h_t \in \mathcal{H}$ and the battery state $b_t \in \mathcal{B}$, i.e., $\boldsymbol{s}_t = [h_t, b_t]$. Similar to the works in [12]–[15], quantized system state is considered in this paper. Specifically, the system state space $\mathcal{S}$ is expressed as $\mathcal{S} = \mathcal{H} \times \mathcal{B}$, where $\mathcal{H} \triangleq \{1, 2, ..., K\}$ and $\mathcal{B} \triangleq \{0, ..., l, ..., L\}$ define the set of channel state and battery state, respectively. The battery is at state 0 when the stored energy is exhausted.

In practice, the channel in a communication system is generally time-correlated. As stated in [15]–[17], the time-varying behavior of the fading channel can be well captured by the FSMC model. Accordingly, in this paper, we separate the channel gain by a set of boundaries, i.e., $\boldsymbol{\Gamma} = \{\Theta_1, \Theta_2, ..., \Theta_k, ...\Theta_{K+1}\} \times d^{-\alpha}$, where Θ_k varies in an increasing order with $\Theta_1 = 0$ and $\Theta_{K+1} = \infty$. In the t-th block, the channel state $h_t \in \mathcal{H}$ is said to be at state k (i.e., $h_t = k$) if $\Theta_k \le \theta_t < \Theta_{k+1}$.

We assume that there is only an one-step channel state transition from block to block. Denoting π_k as the steady state probability of the channel being at state k. With equiprobable partition of the channel gain (this is a reasonable and commonly adopted technique in a FSMC model, cf. [13]–[15]), i.e., $\pi_k = \frac{1}{K}, \forall k \in \{1, 2, ..., K\}$, the fading boundaries Θ_k can be obtained by solving the following equations:

$$\pi_k = \int_{\Theta_k}^{\Theta_{k+1}} \rho(\theta_t) d\theta_t = \frac{1}{K}, \forall k \in \{1, 2, ..., K\}, \qquad (1)$$

where $\rho(\theta_t)$ is the probability density function of the variable θ_t. When channel is at state k, i.e., $h_t = k$, the quantized value of the channel gain is

$$\bar{H}_t = \frac{\int_{\Theta_k}^{\Theta_{k+1}} H_t \rho(\theta_t) \mathrm{d}\theta_t}{\int_{\Theta_k}^{\Theta_{k+1}} \rho(\theta_t) \mathrm{d}\theta_\mathrm{t}} = \frac{\int_{\Theta_k}^{\Theta_{k+1}} \theta_t d^{-\alpha} \rho(\theta_t) \mathrm{d}\theta_t}{\pi_k}. \qquad (2)$$

Similarly, the available energy in the battery of the UE is discretized into L quantum. Denote Q as one energy quantum level of the battery, then the maximum capacity of the battery is $B_{\max} = LQ$. In the t-th block, the battery state is said to be at state l (i.e., $b_t = l$) if $\lfloor \frac{B_t}{Q} \rfloor = l$, where B_t is the available battery energy at the beginning of block t.

B. Actions, Reward, and Cost Functions

At the beginning of each block, the H-AP makes a decision according to the current system state and reports it to the UE such that the system is well scheduled during the following WET and WIT procedure. The time duration of each block T is divided into two orthogonal time slots: τ_t^{E} for WET and τ_t^{I} for WIT with $\tau_t^{\mathrm{E}} + \tau_t^{\mathrm{I}} \le T$. Let P_t^{E} and P_t^{I} be the transmit power of the H-AP for WET and the transmit power of the UE for WIT, respectively. Then, the action adopted in block t (denoted by $\boldsymbol{a}_t$) contains four elements, i.e., $\boldsymbol{a}_t = \{\tau_t^{\mathrm{E}}, \tau_t^{\mathrm{I}}, P_t^{\mathrm{E}}, P_t^{\mathrm{I}}\}$.

For a given system state, different actions come with different rewards and costs. In our work, we consider the throughput per block (defined as the data bits transmitted in one block) as the immediate reward and the energy consumption per block as the immediate cost. Denote the feasible action set at state $\boldsymbol{s}_t$ as $\mathcal{A}(\boldsymbol{s}_t)$. For a given state $\boldsymbol{s}_t$ and an action $\boldsymbol{a}_t \in \mathcal{A}(\boldsymbol{s}_t)$, the immediate reward, i.e., $r(\boldsymbol{s}_t, \boldsymbol{a}_t) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, is defined as

$$r(\boldsymbol{s}_t, \boldsymbol{a}_t) = \frac{\int_{\Theta_k}^{\Theta_{k+1}} \tau_t^{\mathrm{I}} W \log_2 \left(1 + \frac{P_t^{\mathrm{I}} \theta_t d^{-\alpha}}{\zeta \sigma^2}\right) \rho(\theta_t) \mathrm{d}\theta_\mathrm{t}}{\pi_k}, \qquad (3)$$

where W is the bandwidth of the considered system, $\sigma^2 = N_0 W$ is the thermal noise power (where N_0 is the noise power density), and the factor ζ characterizes the discrepancy between the achievable rate and the channel capacity due to the use of practical modulation and coding schemes [4].

The corresponding immediate cost, i.e., $e(\boldsymbol{s}_t, \boldsymbol{a}_t) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, is expressed as

$$e(\boldsymbol{s}_t, \boldsymbol{a}_t) = \frac{P_t^{\mathrm{E}} \tau_t^{\mathrm{E}}}{\vartheta_{\mathrm{AP}}} + P_{\mathrm{C_{AP}}} \tau_t^{\mathrm{E}} + e_t^{\mathrm{IT}} - e_t^{\mathrm{AC}}, \tag{4}$$

where the first two terms capture the energy consumption at the H-AP and the last two terms describe the battery consumption at the UE. Specifically, $0 < \vartheta_{\mathrm{AP}} < 1$ is the power amplifier efficiency of H-AP. Hence the first term in (4) presents the energy consumption of the power amplifier during WET. $P_{\mathrm{C_{AP}}}$ is the circuit power at the H-AP. Hence the second term in (4) accounts for the energy consumption of the circuit during WET. For the battery consumption at the UE,

$$e_t^{\mathrm{IT}} = \frac{P_t^{\mathrm{I}} \tau_t^{\mathrm{I}}}{\vartheta_{\mathrm{U}}} + P_{\mathrm{C_U}} \tau_t^{\mathrm{I}} \tag{5}$$

stands for the energy consumption of the UE during WIT, where ϑ_{U} and $P_{\mathrm{C_U}}$ denote the power amplifier efficiency and the circuit power at the UE, respectively. Finally,

$$e_t^{\mathrm{AC}} = \min\left(B_t + \eta G_{\mathrm{A}} P_t^{\mathrm{E}} \tau_t^{\mathrm{E}} \bar{H}_t, B_{\max}\right) - B_t \tag{6}$$

is the energy accumulated in the battery in block t, where η is the energy conversion efficiency and G_{A} is the antenna gain at the H-AP during WET. Obviously, the value of $e_t^{\mathrm{IT}} - e_t^{\mathrm{AC}}$ can be either positive (battery consumption) or negative (battery accumulation).

By the conservation of energy, both $r(\boldsymbol{s}_t, \boldsymbol{a}_t)$ and $e(\boldsymbol{s}_t, \boldsymbol{a}_t)$ are nonnegative. Since the available energy of the UE in block t is limited by the current stored energy in the battery, the feasible action set at system state $\boldsymbol{s}_t$ can be given as:

$$\begin{aligned}\mathcal{A}(\boldsymbol{s}_t) = &\{\boldsymbol{a}_t | \tau_t^{\mathrm{E}} + \tau_t^{\mathrm{I}} \le T, \tau_t^{\mathrm{E}} \ge 0, \tau_t^{\mathrm{I}} \ge 0, P_t^{\mathrm{I}} \ge 0,\\ &0 \le P_t^{\mathrm{E}} \le P_{\max}^{\mathrm{E}}, e_t^{\mathrm{IT}} \le e_t^{\mathrm{AC}} + B_t\},\end{aligned} \tag{7}$$

where $P_{\max}^{\mathrm{E}}$ is the maximum transmit power of the H-AP.

C. Transition Probabilities

Denote the system state in block t and $t+1$ as $\boldsymbol{s}_t$ and $\boldsymbol{s}_{t+1}$, respectively. For an adopted action $\boldsymbol{a}_t$, the transition probability from state $\boldsymbol{s}_t$ to state $\boldsymbol{s}_{t+1}$ can be given as

$$\begin{aligned}\mathcal{P}(\boldsymbol{s}_{t+1} | \boldsymbol{s}_t, \boldsymbol{a}_t) &\overset{\text{(a)}}{=} \mathcal{P}(h_{t+1}, b_{t+1} | h_t, b_t, \boldsymbol{a}_t)\\ &\overset{\text{(b)}}{=} \mathcal{P}(h_{t+1} | h_t) \mathcal{P}(b_{t+1} | h_t, b_t, \boldsymbol{a}_t),\end{aligned} \tag{8}$$

where (a) holds by definition and (b) holds for the independence of the channel state evolution from the battery state and the action. In the following, we calculate the channel state transition probability $\mathcal{P}(h_{t+1}|h_t)$ and the battery state transition probability $\mathcal{P}(b_{t+1}|h_t, b_t, \boldsymbol{a}_t)$, respectively.

The channel state transition probability, which is closely related to the time-varying behavior of the channel gain, can be described by the level crossing rate $\Lambda(\Theta)$ [15]–[17], i.e., the average number of times that the instantaneous value of θ_t crosses a given level Θ. Specifically, the channel state transition probability from state h_t to h_{t+1} can be approximated by the ratio of $\Lambda(\Theta)$ divided by the average number of blocks the value of θ_t falls in the interval associated with the state h_t. Similar to [13]–[17], we assume that the channel state transits between its adjacent state only (the validity of this commonly-used assumption has been verified in [16]). Then, the channel transition probabilities can be approximated as

$$\mathcal{P}(h_{t+1} = k+1 | h_t = k) \approx \frac{\Lambda(\Theta_{k+1}) T}{\pi_k}, \tag{9}$$

$$\mathcal{P}(h_{t+1} = k-1 | h_t = k) \approx \frac{\Lambda(\Theta_{k-1}) T}{\pi_k}, \tag{10}$$

$$\mathcal{P}(h_{t+1} = k | h_t = k) \approx 1 - \frac{\Lambda(\Theta_{k+1}) T}{\pi_k} - \frac{\Lambda(\Theta_{k-1}) T}{\pi_k}. \tag{11}$$

On the other hand, the battery state transition can be described as follows. If $b_{t+1} < L$,

$$\mathcal{P}(b_{t+1} | h_t, b_t, \boldsymbol{a}_t) = \delta\{b_t + \lfloor \frac{e_t^{\mathrm{AC}} - e_t^{\mathrm{IT}}}{Q} \rfloor = b_{t+1}\}, \tag{12}$$

otherwise,

$$\mathcal{P}(L | h_t, b_t, \boldsymbol{a}_t) = \delta\{b_t + \lfloor \frac{e_t^{\mathrm{AC}} - e_t^{\mathrm{IT}}}{Q} \rfloor \ge L\}, \tag{13}$$

where $\delta(\cdot)$ is the indicator function.

III. CMDP Formulation and The Optimal Policy

In this section, we formulate the CMDP problem and provide the corresponding optimal solution.

A. Problem Formulation

For a system in the long run, a policy $\boldsymbol{\mu}$ is a sequence of decision rules, i.e., $\boldsymbol{\mu} = \{\mu_1, \mu_2, ...\}$, each in which is a function mapping from the system state $\boldsymbol{s}$ to the action to be taken, i.e., $\mu_t : \mathcal{S} \to \mathcal{A}$, $\forall t$. A policy $\boldsymbol{\mu}$ is said to be stationary if the decision rule in it is independent with time, i.e., $\mu_1 = \mu_2 = \cdots$. If a policy is stationary and deterministic, then it is called a pure policy. To model the imperfect operation of the system in Fig. 1, we introduce a factor $\lambda \in [0, 1)$ to capture the probability that the system hardware survives from a operation failure in a block. Correspondingly, as described in [18], for an available stationary policy $\boldsymbol{\mu}$, the long-term throughput of the system can be defined as

$$R(\boldsymbol{s}_0, \boldsymbol{\mu}) = (1-\lambda) \sum_{t=1}^{\infty} \lambda^t \mathbb{E}_{\boldsymbol{s}_0}^{\boldsymbol{\mu}} \{r(\boldsymbol{s}_t, \boldsymbol{a}_t)\}, \tag{14}$$

and the long-term energy cost of the system can be defined as

$$E(\boldsymbol{s}_0, \boldsymbol{\mu}) = (1-\lambda) \sum_{t=1}^{\infty} \lambda^t \mathbb{E}_{\boldsymbol{s}_0}^{\boldsymbol{\mu}} \{e(\boldsymbol{s}_t, \boldsymbol{a}_t)\}. \tag{15}$$

When λ approaches 1, the discounted functions defined in (14) and (15) converge to their expected average values [18], respectively, which are defined in the form of $\lim_{N \to \infty} \frac{1}{N} \sum_{t=1}^{N} \lambda^t \mathbb{E}_{\boldsymbol{s}_0}^{\boldsymbol{\mu}} \{X_t(\boldsymbol{s}_t, \boldsymbol{a}_t)\}, X \in \{r, e\}$, where N is the number of blocks. Thus (14) and (15) can be interpreted as the expected average throughput and the expected average energy cost per block, respectively.

In this paper, we aim at finding an optimal policy $\boldsymbol{\mu}^*$ such that the long-term throughput is maximized under a given energy budget $E_{\rm th}$. This policy can be obtained through solving the following CMDP problem:

$$\max_{\boldsymbol{\mu}} \quad R(\boldsymbol{s}_0, \boldsymbol{\mu}) \tag{16a}$$

$$\text{s.t.} \quad E(\boldsymbol{s}_0, \boldsymbol{\mu}) \leq E_{\rm th}. \tag{16b}$$

B. The Optimal Policy

As shown in [18], the CMDP problem in the form of (16) can be efficiently solved via the Lagrangian approach, whereby the CMDP problem is transferred into an equivalent unconstrained MDP problem. Accordingly, by introducing a non-negative Lagrangian multiplier β for problem (16), a new reward function $\widetilde{r}(\boldsymbol{s}, \boldsymbol{a}; \beta) \ : \ \mathcal{S} \times \mathcal{A} \times \mathbb{R}^+ \rightarrow \mathbb{R}$, can be constructed for the equivalent unconstrained MDP problem, where

$$\widetilde{r}(\boldsymbol{s}, \boldsymbol{a}; \beta) = r(\boldsymbol{s}, \boldsymbol{a}) - \beta e(\boldsymbol{s}, \boldsymbol{a}), \tag{17}$$

and the corresponding Bellman's optimality equation is:

$$J_\beta(\boldsymbol{s}) = \max_{\boldsymbol{a} \in \mathcal{A}(\boldsymbol{s})} \Bigg\{ (1-\lambda)\widetilde{r}(\boldsymbol{s}, \boldsymbol{a}; \beta) + \lambda \sum_{\boldsymbol{s}' \in \mathcal{S}} \mathcal{P}(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a}) J_\beta(\boldsymbol{s}') \Bigg\}, \tag{18}$$

which can be efficiently solved via the Value Iteration Algorithm (VIA) [19] for any fixed β. Correspondingly, the optimal policy with a given β, i.e., $\boldsymbol{\mu}_\beta = \{\mu_\beta(\boldsymbol{s}), \forall \boldsymbol{s} \in \mathcal{S}\}$, can be determined by:

$$\mu_\beta(\boldsymbol{s}) = \arg \max_{\boldsymbol{a} \in \mathcal{A}(\boldsymbol{s})} \Bigg\{ (1-\lambda)\widetilde{r}(\boldsymbol{s}, \boldsymbol{a}; \beta) + \lambda \sum_{\boldsymbol{s}' \in \mathcal{S}} \mathcal{P}(\boldsymbol{s}'|\boldsymbol{s}, \boldsymbol{a}) J_\beta(\boldsymbol{s}') \Bigg\}. \tag{19}$$

As described in [18], the optimal policy of a CMDP problem with a single constraint is composed of two pure policies, i.e., $\boldsymbol{\mu}_{\beta^-}$ and $\boldsymbol{\mu}_{\beta^+}$, with β^- and β^+ as their associated Lagrangian multipliers, respectively. The policy $\boldsymbol{\mu}_{\beta^-}$ yields the highest energy cost E^- that satisfies the energy constraint, while the policy $\boldsymbol{\mu}_{\beta^+}$ yields the lowest energy cost E^+ that breaks the energy constraint. Since $J_\beta(\boldsymbol{s})$ is a monotonically non-increasing function of β [20], the value of β^- and β^+ can be efficiently obtained via the bisection search method. With a randomized mixture of $\boldsymbol{\mu}_{\beta^-}$ and $\boldsymbol{\mu}_{\beta^+}$, the optimal policy of a CMDP problem can be given by:

$$\boldsymbol{\mu}^* = \begin{cases} \boldsymbol{\mu}_{\beta^-}, & \text{w.p. } q \qquad (20) \\ \boldsymbol{\mu}_{\beta^+}, & \text{w.p. } 1-q \ , \qquad (21) \end{cases}$$

where the mixing weight parameter $0 \leq q \leq 1$ can be obtained via solving equation $E_{\rm th} = qE^- + (1-q)E^+$.

Correspondingly, the procedures for solving problem (16) is described in Algorithm 1. Since the optimal policy consists of two pure policies, both of which are irrelevant to time sequence. In Algorithm 1, we drop the subscript "t" for convenience. Specifically, initializations are performed in line 1, where n and ε_β are the iteration sequence and the error bound for updating β, respectively. The initial value of β^+ is specified in an incremental method, i.e., increasing the initial value of β^+ until that the corresponding long-term system energy cost exceeds $E_{\rm th}$. The VIA is conducted to solve the equivalent unconstrained MDP problem in line 4 and the Lagrangian multiplier β is updated via bisection search in lines 5-13. Finally, with the obtained policy $\mu_{\beta^-}(\boldsymbol{s})$ and $\mu_{\beta^+}(\boldsymbol{s})$, the mixing weight q and the optimal policy are obtained in line 17 and line 18, respectively.

For the implementation of VIA, the candidate actions at each state are quantized. Specifically, $\tau^{\rm E}$, $\tau^{\rm I}$, $P^{\rm E}$, and $P^{\rm I}$ are discretized into levels of $V_\tau^{\rm E}$, $V_\tau^{\rm I}$, $V_P^{\rm E}$, and $V_P^{\rm I}$, respectively. Since the update of β is independent from the action space and the channel state space, the computational complexity of Algorithm 1 is $\mathcal{O}(\frac{1}{1-\lambda}\log(\frac{1}{1-\lambda})V_\tau^{\rm E}V_\tau^{\rm I}V_P^{\rm E}V_P^{\rm I}|\mathcal{S}|^3)$ [21].

Remark 1: In this paper, we obtain the optimal online policy for the CMDP problem (16) for the case of single UE. For the case of $M > 1$ UEs, the corresponding tuple of the CMDP can be constructed as follow (here, we use the subscript "m" to denote the elements of the m-th UE): the system space $\bar{\mathcal{S}}$ can be expressed as $\bar{\mathcal{S}} = \mathcal{S}_1 \times \mathcal{S}_2 ... \times \mathcal{S}_m ... \times \mathcal{S}_M$, where $\mathcal{S}_m = \mathcal{H}_m \times \mathcal{B}_m$ is the system state space of the m-th UE and "$\times$" is the Cartesian product; the action space $\bar{\mathcal{A}}$ can be expressed as $\bar{\mathcal{A}} = \mathcal{A}_1 \times ...\mathcal{A}_m ... \times \mathcal{A}_M$, where $\mathcal{A}_m$ presents the action space of the m-th UE and is in the form of (7); for an action $\bar{\boldsymbol{a}}_t = [\boldsymbol{a}_{1,t}, ..., \boldsymbol{a}_{m,t}, ..., \boldsymbol{a}_{M,t}]$ adopted at state $\bar{\boldsymbol{s}}_t = [\boldsymbol{s}_{1,t}, ..., \boldsymbol{s}_{m,t}, ..., \boldsymbol{s}_{M,t}]$, the immediate reward and the immediate cost of the system can be defined as $\bar{r}(\bar{\boldsymbol{s}}_t, \bar{\boldsymbol{a}}_t) = \sum_{m=1}^{M} r(\boldsymbol{s}_{m,t}, \boldsymbol{a}_{m,t})$ and $\bar{e}(\bar{\boldsymbol{s}}_t, \bar{\boldsymbol{a}}_t) = \sum_{m=1}^{M} e(\boldsymbol{s}_{m,t}, \boldsymbol{a}_{m,t})$, respectively; the system state transition probability matrix can be expressed as $\mathbb{P} = \mathbb{P}_1 \otimes ...\mathbb{P}_m ... \otimes \mathbb{P}_M$, where $\mathbb{P}_m = [\mathcal{P}(\boldsymbol{s}_{m,t+1}|\boldsymbol{s}_{m,t}, \boldsymbol{a}_{m,t})]$ is the system state transition probability matrix of the m-th UE and $\otimes$ is the Kronecker product. Based on this tuple, the CMDP problem for the multi-user case can be constructed and the corresponding optimal online policy can be obtained similarly via Algorithm 1.

Algorithm 1 The Optimal Policy for the CMDP (16)

```
1:  Set $n = 0$, $\beta^- = 0$, $\beta^+$, $\beta^0 = \beta^-$, specify $\varepsilon_\beta > 0$.
2:  repeat
3:      Set $\beta = \beta^n$ and $n = n + 1$.
4:      For a given $\beta$, obtain the optimal policy $\boldsymbol{\mu}_\beta = \{\mu_\beta(\boldsymbol{s}), \forall \boldsymbol{s} \in \mathcal{S}\}$ via VIA.
5:      Compute the stationary distribution $\Psi(\boldsymbol{s})$ induced by $\boldsymbol{\mu}_\beta = \{\mu_\beta(\boldsymbol{s}), \forall \boldsymbol{s} \in \mathcal{S}\}$.
6:      if $\sum_{\boldsymbol{s} \in \mathcal{S}} \Psi(\boldsymbol{s}) e(\boldsymbol{s}, \mu_\beta(\boldsymbol{s})) > E_{\rm th}$ then
7:          $\beta^{n+1} = \frac{\beta^+ + \beta^n}{2}$.
8:          $\beta^- = \beta^n$.
9:      else
10:         $\beta^{n+1} = \frac{\beta^- + \beta^n}{2}$.
11:         $\beta^+ = \beta^n$.
12:     end if
13: until $|\beta^{n+1} - \beta^n| < \varepsilon_\beta$.
14: Find the policies $\boldsymbol{\mu}_{\beta^-} = \{\mu_{\beta^-}(\boldsymbol{s}), \forall \boldsymbol{s} \in \mathcal{S}\}$ and $\boldsymbol{\mu}_{\beta^+} =$
```

```
    $\{\mu_{\beta^+}(\boldsymbol{s}), \forall \boldsymbol{s} \in \mathcal{S}\}$ with obtained $\beta^-$ and $\beta^+$, respectively.
Compute the stationary distribution $\Psi_{\beta^-}(\boldsymbol{s})$ and $\Psi_{\beta^+}(\boldsymbol{s})$ induced by $\boldsymbol{\mu}_{\beta^-}$ and $\boldsymbol{\mu}_{\beta^+}$, respectively.
Compute
    $R_{\beta^-} = \sum_{\boldsymbol{s} \in \mathcal{S}} \Psi_{\beta^-}(\boldsymbol{s}) r(\boldsymbol{s}, \mu_{\beta^-}(\boldsymbol{s})), \tag{22}$
    $R_{\beta^+} = \sum_{\boldsymbol{s} \in \mathcal{S}} \Psi_{\beta^+}(\boldsymbol{s}) r(\boldsymbol{s}, \mu_{\beta^+}(\boldsymbol{s})), \tag{23}$
    $E_{\beta^-} = \sum_{\boldsymbol{s} \in \mathcal{S}} \Psi_{\beta^-}(\boldsymbol{s}) e(\boldsymbol{s}, \mu_{\beta^-}(\boldsymbol{s})), \tag{24}$
    $E_{\beta^-} = \sum_{\boldsymbol{s} \in \mathcal{S}} \Psi_{\beta^+}(\boldsymbol{s}) e(\boldsymbol{s}, \mu_{\beta^+}(\boldsymbol{s})). \tag{25}$
Compute $q$ by solving $E_{\text{th}} = qE_{\beta^-} + (1-q)E_{\beta^+}$.
Obtian the optimal reward $R = qR_{\beta^-} + (1-q)R_{\beta^+}$ and the optimal policy
    $\boldsymbol{\mu}^* = \begin{cases} \boldsymbol{\mu}_{\beta^-}, & \text{w.p. } q \quad (26) \\ \boldsymbol{\mu}_{\beta^+}, & \text{w.p. } 1-q \quad (27) \end{cases}$
```

IV. Simulation Results

In this section, numerical simulations are provided for evaluating the long-term throughput performance of the system. For the practicality of RF energy transfer, a Rician fading channel is considered between the H-AP and the UE [22], [23]. Correspondingly, the PDF of θ_t is given by

$$\rho(\theta_t) = \frac{1}{2\varrho^2} e^{\frac{-(\theta_t + \varsigma^2)}{2\varrho^2}} I_0 \left(\frac{\sqrt{\theta}_t \varsigma}{\varrho^2} \right), \tag{28}$$

where I_0 is the modified Bessel function of the zero-th order, $2\varrho^2$ and ς^2 are the parameters representing the power of multi-path and line-of-sight, respectively. Moreover, the level crossing rate $\Lambda(\Theta_{\text{b}})$ is [17]

$$\Lambda(\Theta) = \sqrt{\frac{2\pi(1+\kappa)\Theta}{\bar{\theta}}} f_D e^{-(\kappa + \frac{1+\kappa}{\bar{\theta}}\Theta)} I_0 (2\sqrt{\frac{\kappa(1+\kappa)\Theta}{\bar{\theta}}}), \tag{29}$$

where f_D is the maximum Doppler shift of the channel, $\bar{\theta} = 2\varrho^2 + \xi^2$ is the local-mean fading power and $\kappa = \frac{\xi^2}{2\varrho^2}$. Accordingly, practical channel parameters setting in [17] is considered in simulations, where the number of channel states is selected as $K = 3$, f_D is set as 1.34 Hz, and the block duration is set as $T = 16$ ms, respectively.

Similar to [10], we focus on the case of small devices and express the battery size as a function of the reference value $B_{\text{ref}} = 10^{-3} \times T$ J. Unless otherwise stated, the maximum battery capacity is set as $B_{\max} = 10B_{\text{ref}}$. On the other hand, extensive simulations (not shown here) have revealed that the accuracy of results is guaranteed when $\varepsilon_\beta = 10^{-4}$ and $Q = B_{\text{ref}}$. Other important parameters used in simulations are listed in Table I. Moreover, to show the superiority of the optimal policy, the myopic policy which maximizes the throughput in only the current block is used as the benchmark. For legibility, in the simulation results, we mark the optimal policy and the myopic policy as "Optimal" and "Myopic", respectively.

TABLE I
Parameters Setting

$P^{\text{E}}_{\max}$	10 W	α	2.8	$P_{\text{C}_{\text{AP}}}$	500 mW
$P_{\text{C}_{\text{U}}}$	5 mW	ϑ_{AP}	0.9	ϑ_{U}	0.9
η	0.95	λ	0.9	G_A	8 dBi
ζ	1	W	2 kHz	N_0	-164 dBm/Hz
ς^2	0.75	ϱ^2	0.125		

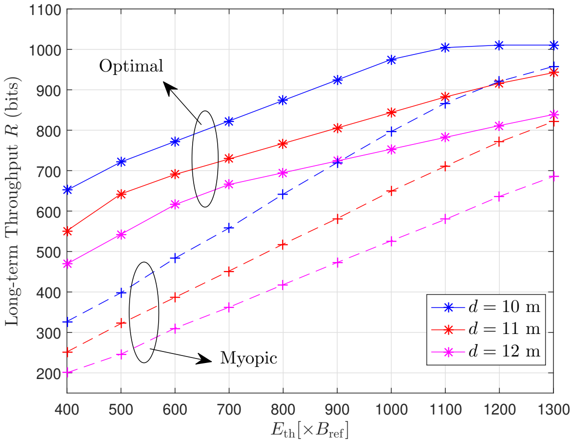

Fig. 2. The long-term throughput versus the system energy budget E_{th}.

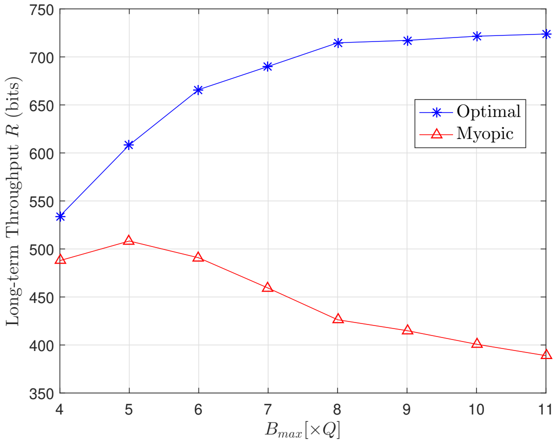

Fig. 3. The long-term throughput versus the maximum battery capacity $B_{\max}$.

To investigate the impact of the energy budget and the communication distance on the system throughput performance, we first depict the long-term throughput as as a function of the energy budget E_{th} for different value of d. As shown in Fig. 2, the optimal policy outperforms the myopic policy in all the considered cases. The long-term throughput is shown to be increased with E_{th}. This is because that a larger E_{th} means more available energy budget. Due to the limitation of transmit power and the battery capacity, the system performance becomes saturated when E_{th} is exceedingly large (see the case of d=10 m). On the other hand, since the signal attenuations during WIT and WET are decreasing functions of the communication distance. As expected, the long-term throughput is shown to be reduced with d.

The maximum battery capacity $B_{\max}$, which limits the maximum available energy at the UE in each block, is expected to have an impact on the system performance. Hence, in Fig. 3, we investigate the long-term system throughput with varying $B_{\max}$. Here, we set $E_{\text{th}} = 500B_{\text{ref}}$ and $d = 10$ m. As shown in the figure, the long-term throughput with the optimal policy increases with $B_{\max}$. In fact, a larger $B_{\max}$ means a higher ability to handle the fluctuation of the channel state. As $B_{\max}$ grows, the performance gain becomes saturate due to the limitation of E_{th}. However, the myopic policy shows a different trend. With the growth of $B_{\max}$, the corresponding long-term throughput first increases and then decreases. This is due to the fact that the myopic policy operates sequentially from block to block and exhausts the battery's energy as much as possible to maximize the current system throughput, which results in a trade-off on $B_{\max}$. Nevertheless, compared to the myopic policy, considerable improvement can be observed when the optimal policy is adopted.

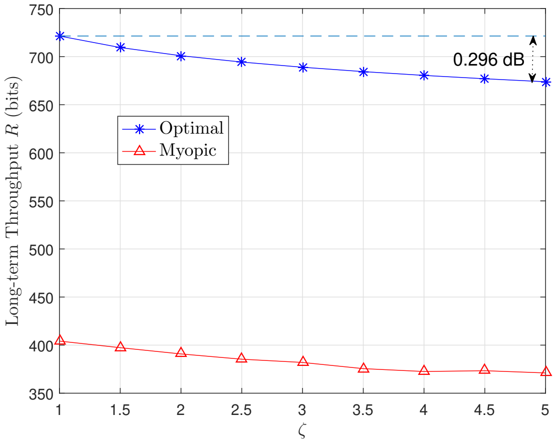

Fig. 4. The long-term throughput versus the gap factor ζ.

As stated in (3), the factor ζ is used to capture the impact from the practical modulation and coding schemes. In Fig. 4, we depict the long-term throughput as a function of ζ with $E_{\text{th}} = 500B_{\text{ref}}$ and $d = 10$ m. As shown in the figure, compared with the myopic policy, a high system performance gain is achieved when the optimal policy is adopted. Moreover, the long-term throughput is shown to be slightly decreased with the increasing ζ. For example, with rising ζ from 1 to 5 (about 7 dB), the long-term throughput performance for the optimal policy drops only about 0.296 dB. On the other hand, the impact of ζ on the optimal policy is investigated in Fig. 5. Here, we take the optimal policy with $\zeta = 1$ (i.e., $\boldsymbol{\mu}^*_{\zeta=1}$) as the reference policy and use a binary indicator C to identify the variation of the optimal policy with ζ. Specifically, denote the optimal policy with ζ' as $\boldsymbol{\mu}^*_{\zeta'}$, then $C = 1$ if $\boldsymbol{\mu}^*_{\zeta'}$ is identical to $\boldsymbol{\mu}^*_{\zeta=1}$. Otherwise, $C = 0$. As demonstrated in Fig. 5, the value of C equals to 1 and remains unchanged for different values of ζ, which implies that the optimal policy is irrelevant to the practical implementation of the modulation and coding schemes.

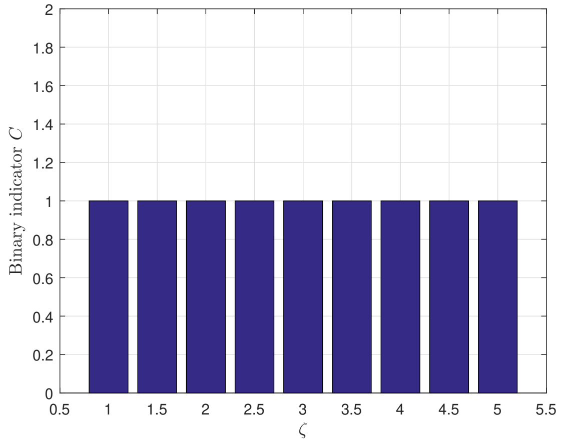

Fig. 5. The binary indicator C versus the gap factor ζ.

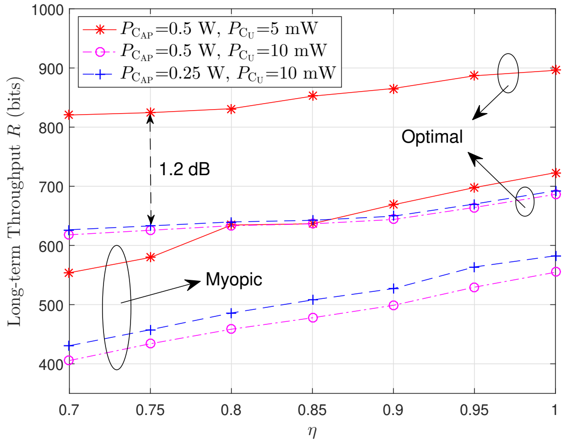

Fig. 6. The long-term throughput versus the energy conversion efficiency η with different circuit power P_{CAP} and P_{CU}.

Lastly, the impact of the energy conversion efficiency and the circuit power on the system performance is investigated in Fig. 6. Here we set $E_{\text{th}} = 500B_{\text{ref}}$ and $d = 8$m. As can be observed, the long-term throughput grows with the increasing of η. This is due to the fact that more available energy can be harvested at the UE with higher energy conversion efficiency. On the other hand, although P_{CAP} dominates the circuit power of the whole system, the system performance is shown to be more sensitive to P_{CU} rather than P_{CAP}. Specifically, with the optimal policy, the long-term throughput achieves a performance gain of 1.2 dB at $\eta = 0.75$ when P_{CU} decreases 3 dB (from 10 mW to 5 mW), but is almost unchanged when P_{CAP} drops from 0.5 W to 0.25 W. In practice, this intrigues an prior effort on cutting down the circuit power consumption at the UE rather than at the H-AP.

V. Conclusion

In this paper, we studied the problem of designing the optimal online policy in an energy-constrained WPCN to manage the transmit power and time durations for both WET and WIT over time-correlated fading channels. Aiming at maximizing the system long-term throughput with a limited energy budget, we formulate the transmission policy design as a CMDP problem, which was later transformed into an equivalent unconstrained MDP problem and solved via the Lagrangian approach. Numerical results showed that the long-term system performance is closely related to the total energy budget, the battery capacity, the communication distance, the energy conversion efficiency, and the circuit power of the system. For instance, the circuit power consumption at the UE has a stronger impact on the system performance than that at the H-AP. Also, the optimal policy was shown to be independent of the choices of modulation and coding schemes.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (U1713209, 61520106009, 61533008, 61573103) and in part by CSC. The work of X. Zhou was supported by the Australian Research Council's Discovery Projects Funding Scheme (Project number DP170100939). The work of D. W. K. Ng was supported by the Australian Research Councils Discovery Early Career Researcher Award (DE170100137).

References

[1] Q. Wu, G. Y. Li, W. Chen, D. W. K. Ng, and R. Schober, "An overview of sustainable green 5G networks," *IEEE Wireless Commun.*, vol. 24, no. 4, pp. 72–80, Aug. 2017.

[2] H. Ju and R. Zhang, "Throughput maximization in wireless powered communication networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 1, pp. 418–428, Jan. 2014.

[3] H. Chen, Y. Li, J. L. Rebelatto, B. F. Ucha-Filho, and B. Vucetic, "Harvest-then-cooperate: Wireless-powered cooperative communications," *IEEE Trans. Signal Process.*, vol. 63, no. 7, pp. 1700–1711, Apr. 2015.

[4] Q. Wu, M. Tao, D. W. K. Ng, W. Chen, and R. Schober, "Energy-efficient resource allocation for wireless powered communication networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 2312–2327, Mar. 2016.

[5] H. Kim, H. Lee, M. Ahn, H. Kong, and I. Lee, "Joint subcarrier and power allocation methods in full duplex wireless powered communication networks for OFDM systems," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4745–4753, Jul. 2016.

[6] S. Luo, G. Yang, and K. C. Teh, "Throughput of wireless-powered relaying systems with buffer-aided hybrid relay," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4790–4801, Jul. 2016.

[7] P. D. Diamantoulakis, K. N. Pappi, Z. Ding, and G. K. Karagiannidis, "Optimal design of non-orthogonal multiple access with wireless power transfer," in *Proc. IEEE Int. Conf. Communications (ICC)*, May 2016, pp. 1–6.

[8] R. Morsi, D. S. Michalopoulos, and R. Schober, "Performance analysis of near-optimal energy buffer aided wireless powered communication," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 863–881, Feb. 2018.

[9] X. Zhou, C. K. Ho, and R. Zhang, "Wireless power meets energy harvesting: A joint energy allocation approach in OFDM-based system," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3481–3491, May 2016.

[10] A. Biason and M. Zorzi, "Battery-powered devices in WPCNs," *IEEE Trans. Commun.*, vol. 65, no. 1, pp. 216–229, Jan. 2017.

[11] M. A. Abd-Elmagid, A. Biason, T. ElBatt, K. G. Seddik, and M. Zorzi, "On optimal policies in full-duplex wireless powered communication networks," in *Proc. 14th Int. Symp. Modeling and Optimization in Mobile, Ad Hoc and Wireless Networks (WiOpt)*, May 2016, pp. 1–7.

[12] S. Mao, M. H. Cheung, and V. W. S. Wong, "Joint energy allocation for sensing and transmission in rechargeable wireless sensor networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 6, pp. 2862–2875, Jul. 2014.

[13] B. Li, W. Guo, Y. Liang, C. An, and C. Zhao, "Asynchronous device detection for cognitive device-to-device communications," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2443–2456, Apr. 2018.

[14] R. Zhang, Z. Zhong, Y. Zhang, S. Lu, and L. Cai, "Measurement and analytical study of the correlation properties of subchannel fading for noncontiguous carrier aggregation," *IEEE Trans. Veh. Technol.*, vol. 63, no. 9, pp. 4165–4177, Nov. 2014.

[15] P. Sadeghi, R. A. Kennedy, P. B. Rapajic, and R. Shams, "Finite-state Markov modeling of fading channels - a survey of principles and applications," *IEEE Signal Process. Mag.*, vol. 25, no. 5, pp. 57–80, Sep. 2008.

[16] H. S. Wang and N. Moayeri, "Finite-state Markov channel-a useful model for radio communication channels," *IEEE Trans. Veh. Technol.*, vol. 44, no. 1, pp. 163–171, Feb. 1995.

[17] F. Babich and G. Lombardi, "A Markov model for the mobile propagation channel," *IEEE Trans. Veh. Technol.*, vol. 49, no. 1, pp. 63–73, Jan. 2000.

[18] E. Altman, *Constrained Markov decision processes*. Chapman & Hall/CRC, 1998.

[19] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, NJ, USA: Wiley, 2005.

[20] F. J. Beutler and K. W. Ross, "Optimal policies for controlled Markov chains with a constraint," *Journal of Mathematical Analysis and Applications*, vol. 112, pp. 236–252, Nov. 1985.

[21] M. L. Littman, T. L. Dean, and L. P. Kaelbling, "On the complexity of solving Markov decision problems," in *Proc. the Eleventh Conf. Uncertainty in artificial intelligence - UAI '95*, Aug. 1995, pp. 394–402.

[22] Y. Zeng and R. Zhang, "Optimized training design for wireless energy transfer," *IEEE Trans. Commun.*, vol. 63, no. 2, pp. 536–550, Feb. 2015.

[23] F. Zhao, H. Lin, C. Zhong, Z. Hadzi-Velkov, G. K. Karagiannidis, and Z. Zhang, "On the capacity of wireless powered communication systems over Rician fading channels," *IEEE Trans. Commun.*, vol. 66, no. 1, pp. 404–417, Jan. 2018.