IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 23, 2019, accepted September 10, 2019, date of publication September 13, 2019,
date of current version September 25, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2941377

Optimal Finite Horizon Sensing for Wirelessly Powered Devices

MEHDI SALEHI HEYDAR ABAD AND OZGUR ERCETIN
Faculty of Engineering and Natural Sciences, Sabanci University, 34956 Istanbul, Turkey
Corresponding author: Mehdi Salehi Heydar Abad (mehdis@sabanciuniv.edu)

This work was supported in part by the European Commission (EC) H2020-MSCA-RISE-2015 Programme under Grant 690893.

ABSTRACT We are witnessing a significant advancements in the sensor technologies which has enabled a broad spectrum of applications. Often, the resolution of the produced data by the sensors significantly affects the output quality of an application. We study a sensing resolution optimization problem for a wireless powered device (WPD) that is powered by wireless power transfer (WPT) from an access point (AP). We study a class of harvest-first-transmit-later type of WPT policy, where an access point (AP) first employs RF power to recharge the WPD in the down-link, and then, collects the data from the WPD in the up-link. The WPD optimizes the sensing resolution, WPT duration and dynamic power control in the up-link to maximize an application dependant utility at the AP. The utility of a transmitted packet is only achieved if the data is delivered successfully within a finite time. Thus, we first study a finite horizon throughput maximization problem by jointly optimizing the WPT duration and power control. We prove that the optimal WPT duration obeys a time-dependent threshold form depending on the energy state of the WPD. In the subsequent data transmission stage, the optimal transmit power allocations for the WPD is shown to posses a channel-dependent fractional structure. Then, we optimize the sensing resolution of the WPD by using a Bayesian inference based multi armed bandit problem with fast convergence property to strike a balance between the quality of the sensed data and the probability of successfully delivering it.

INDEX TERMS Bayesian inference, multi-armed bandit, reinforcement learning, wireless power transfer.

I. INTRODUCTION

A. MOTIVATION

With the rapid increase in the number of battery-powered devices, energy harvesting (EH) technology provides a convenient window of opportunity to bypass the challenging, and in some cases infeasible task of replacing batteries. Traditional approaches in EH technologies harvest energy from natural resources such as wind, solar, etc. The inherent challenge of EH from natural resources is the stochastic nature of the EH process, which dictates the amount and availability of harvested energy that is beyond the control of system designers. Towards this end, wireless power transfer (WPT) [1] is considered as a promising technology to provide the network administrators a leverage on replenishing the remote devices for proper network operations, by utilizing the RF signals as a mean to transfer power to wireless powered devices (WPDs).

The associate editor coordinating the review of this manuscript and approving it for publication was Yiyu Shi.

WPT brings forth a new dimension of optimization of the performance of sensor networks. In [2], a poll based medium access protocol (MAC) is proposed to collaboratively aide the energy request messages of those sensors that are low on energy. In [3], multiple sensors aim to estimate a parameter of interest in a distributed manner while an Access Point (AP) optimizes the WPT strategy in order to minimize the mean-square error (MSE). In [4], power-splitting and time-splitting schemes utilized in simultaneous wireless information and power transfer (SWIPT) are optimized to maximize the throughput of multiple wireless sensors. In [5], a feasibility analysis of wireless powered sensors under various scenarios is studied to ensure the reliability of energy autonomous critical infrastructure monitoring applications.

WPDs are utilized mainly for collecting and transmitting information for further processing to data collecting units. Traditionally, the scope for the application of sensors were limited to sensing and transmitting fixed-size data packets such as the information regarding temperature, humidity and etc. With the rapid development of hardware technologies for sensors many emerging applications require the transmission

of a much broader type of information. On-body sensors and wearables are examples of these applications where audio, video and gesture information are captured and transmitted to an AP for further processing. The processing includes but not limited to audio, image and video where the resolution of the data points is an important factor in determining the quality of an output produced by an application at hand. For example, the WPD could be an image sensor that transmits images to the AP, tracking the eye movement, i.e., estimating the gaze location of a person [6]. The accuracy of estimating the gaze depends on the number of pixels per frame. A gaze error varies from $10-15$ pixels at 77 pixels/frame to $0-3$ pixels at 1984 pixels/frame [7]. Hence, high resolution sensing provides a better utility in the application layer. However, high resolution sensing compromises the performance of the WPD in two main aspects; first, a high-resolution sensing typically consumes more energy. Second, it generates more data bits per sensing event which may then increase the packet drop probability. Our main objective is to strike a balance between the utility achieved by a sensing configuration and the probability of successfully delivering the sensed packet to the AP.

Optimizing the sensing resolution efficiently requires first addressing the design of WPT scenario. In wireless powered communication networks (WPCNs) [8]–[10], WPT occurs in the down-link (DL) to replenish the battery of WPDs which in turn is used for information transmission (IT) in the up-link (UL). A fundamental question inherited in WPCNs is the optimum duration for WPT period and power allocation in the IT period. We consider a delay sensitive sensing application scenario where the sensed packet needs to be delivered to the AP with a delay that cannot be tolerated beyond the duration of a finite horizon window. The term finite horizon corresponds to a maximum tolerable delay for the involved application. References [8]–[10] perform a single-time-slot optimization assuming that the channel stays constant and all the harvested energy in a slot is totally used in the same time slot. Differently, [11] assumes an infinite horizon throughput maximization problem where the harvested energy is allowed to be used in later times. It was shown that this strategy significantly improves the throughput albeit having high computational complexity.

In the aforementioned works, it is assumed that in a single WPT instance, i.e., transmission of energy in the DL and reception of information in the UL, the channel state stays constant. However, in practice, this assumption is usually not valid, for example due to the body blocking the wearable sensors. In this work, we aim to optimize the sensing resolution of the WPD while jointly optimizing the WPT duration and power allocation in the IT period to maximize the chance of delivering the sensed packet by the WPD to the AP. Particularly, we first study the sub-problem of finite horizon throughput maximization, where both WPT and IT period is exposed to multiple random realizations of channel. The objective is to judiciously determine the optimal WPT duration and power allocations in the IT period. Throughput maximization problem maximizes the chance of delivering the sensed data to AP allowing to simplify the sensing optimization problem. The CSI is available causally and only in the IT period. The availability of causal CSI, makes the problem investigated here challenging, since any decision at any time slot has a cascading effect on the future outcomes.

For the throughput maximization problem, we study the problem under both offline and online settings. In the offline case, CSI is available to the WPD prior to transmission. In other words, at $t = 1$, the WPD knows the CSI for $t = 1, \ldots, T$. In the online case, CSI is available only causally, i.e., the WPD only knows CSI for time t and not for any future time instants. For the offline case, we obtain closed form expressions to find the optimal WPT duration and power allocation in the IT period. We use the insights gained from the offline case, to develop an optimal online policy that maximizes the expected finite horizon throughput by optimally determining the WPT duration and power allocation in the IT period. Specifically, we formulate the problem of optimal WPT duration using the theory of stopping times. A stopping time is a random variable whose value maximizes a certain property of interest in a stochastic process. We show that there exist a time-dependent threshold on the energy level of the WPD in which it is optimal to stop WPT and start the IT period. Then, we show that the optimal power allocation in the IT period follows a fractional structure in which the WPD at each time slot allocates a fraction of its energy that depends on the current channel state as well as a specific measure of future channel expectations.

The optimal policy for determining the WPT period and power allocations in the IT period is used by the WPD to maximize its chance of delivering the sensed packet to the AP for gaining the application specific utility. Hence, as the last part of the solution, we aim to provide a framework where the WPD is able to determine the sensing resolution of the data to be sent to the AP for further processing. A high resolution data increases the performance of the application at the AP; however, a high resolution data has more bits compared to a lower resolution data which may compromise the probability of successfully delivering the data. Therefore, an optimal sensing resolution is required to balance the quality of the sensed data and the probability of successfully delivering it. Due to the dynamic and online nature of the problem, i.e., availability of only causal information, instead of conventional optimization methods, we use Bayesian inference as a reinforcement learning method to provide a mean for the WPD in learning to balance the sensing resolution. We illustrate the benefits of the Bayesian inference over the traditional approaches such as ϵ-greedy algorithm using numerical evaluations.

B. CONTRIBUTIONS

The contributions of the paper are summarized as follows:

- We formulate the problem of finite horizon sensing utility optimization for a WPD. The optimization problem

is first addressed by maximizing the throughput of the WPD and then optimizing the sensing resolution of the sensed data.
- To maximize the throughput, we study the optimization of WPT duration and dynamic power allocation in offline and online settings.
- For the offline problem, where CSI is known non-causally, we derive a closed form expressions that enable a tractable framework to optimize both the WPT duration and power allocation in the IT period. We show that the optimal power allocation has a fractional structure depending on the current channel state as well as future channel states.
- Motivated by the results obtained from the offline problem, we formulate the online problem by assuming that the CSI is available only causally.
- We show that the optimal WPT duration for the online case has a time dependent threshold structure on the available energy of the WPD. We provide an easy to implement method to numerically calculate the thresholds.
- Similar to the offline case, we show that the optimal power allocation for the online counterpart also follows a fractional structure. The WPD allocates a fraction of its available energy in each time slot. Unlike the offline case, optimal fractions in the online case depends on the current channel state and a measure of the future channel state expectations.
- After developing an algorithm capable of maximizing the packet delivery chance, we then focus on optimizing the sensing resolution to maximize a given utility. We employ Bayesian framework based multi-armed bandit problem to learn to determine the resolution of the sensing to balance the quality of the sensed data and the probability of successfully delivering it. We show that the Bayesian framework converges much faster, by judiciously exploring in the action space of the problem, than its classic counterpart ϵ-greedy algorithm.

C. RELATED WORK

WPCN has been studied in the literature under different settings. Reference [14] studies a heterogeneous WPCN with the presence of EH and non-EH devices to find out how the presence of non-harvesting nodes can be utilized to enhance the network performance, compared to pure WPCNs. In [15], problem of throughput maximization in the presence of an EH relay is studied where the relay cooperatively help the source node in relaying its messages. The outage problem for a three node WPCN is analyzed in [16], [17] where both source and relay harvest energy for a certain duration, and then the source transmits to destination by using the relay. Approximate closed-form expressions for outage probability and ergodic capacity in a SWIPT scenario for multiple deployed sensors in [4]. In [46], for a multiuser orthogonal frequency division multiple (OFDM) system employing SWIPT, power-splitting and time splitting modes along with the allocation of the subcarriers are optimized so that the average outage across all users are minimized. Aforementioned works assume a known and time-invariant channel which is unlike our case where we consider a time varying channel with causal CSI. User cooperation is also studied in multiple works [9], [19], [20] to improve the performance of the WPCN by exploiting the cooperative diversity. Multiple works also studied the WPCN in the context of cloud computing [21]–[24]. Throughput maximization for WPCN is studied in [8], [11]–[13]. Per time slot throughput maximization is studied in [8]. By allowing the storage of the energy in a battery by the WPD, [11] studies infinite horizon throughput maximization in HD mode and the results are extended to FD mode in [12]. By adopting a NOMA strategy and under non-causal CSI, [13] studies the problem of finite horizon throughput maximization.

Finite horizon throughput maximization has been extensively addressed from a communication perspective in the literature for non-RF EH techniques. For example, [25] aims at maximizing the finite horizon throughput by dynamically adjusting the transmission power in an offline setting where CSI and the EH information (EHI) is non-causally available at the transmitter for the duration of the deadline. Packet transmission time minimization over a finite horizon with non-causal EHI and a static channel is studied in [26]. However, in practice, the finite horizon spans over multiple time slots, and the CSI and EHI are not usually available. For time varying scenarios where EHI or CSI (or both) are available only causally, the problem needs to be solved dynamically. In [27]–[30] under different EHI and CSI assumptions, the problem of finite horizon throughput maximization is formulated as a dynamic program (DP) and the optimal policy is evaluated by numerically solving the DP. The solution is later stored in the devices as a look-up table. However, the DP solutions are computationally expensive, and they require large memory space to store the solutions, which is usually prohibitive for resource-constrained IoT devices. Moreover, calculating and disseminating the optimal look up tables in a network consisting of large number of WPDs is inherently challenging and introduces large overheads [31]. Finally, the complexity of the numerical solutions increase exponentially with respect to the number of states in the DP formulation. A common way in dealing with such complexity is to reduce the size of the state space (action) of the problem by gaining insight into the dynamic problem [32], [33]. For example, in [32], we have shown that the policy maximizing the infinite horizon throughput of an EH transmitter over a correlated wireless channel exhibits a battery-dependant threshold type structure on the quality of the channel state. In [33], we show that the optimal policy, for an EH receiver equipped with Hybrid ARQ with incremental redundancy (INR), minimizing the expected number of re-transmissions, never splits the incoming RF signal and uses it either for harvesting energy or extracting information. Hence, we convert a continuous state and action Markov decision process (MDP) into a discrete one. Recently, [34] studied the problem of energy efficient scheduling for a non-RF EH over a

finite horizon by developing a low complexity online heuristic policy that is built upon the offline solution and it can achieve close performance with respect to the offline policy. However, albeit the good performance, it is not evident how the algorithm would incorporate the optimal duration of the WPT period. Finally, in [35], we addressed the optimization of the WPT duration and power allocation under a simplified model. Unlike [35], here, we derive an optimal upper bound on the performance of the WPD in terms of the expected throughput over the finite horizon. We extend our results to incorporate a smart sensing application in the WPT scenario where we balance the quality of the sensed data and the probability of successful transmission using reinforcement learning. In [46], WPT is used as an incentive for motivating user involvement in a mobile crowd sensing scenario, where the users store a fraction of the received power as reward and use the rest to sense, compress and transmit a packet back to the AP for maximizing data utility. However, the optimization problem is formulated for a single time slot with constant channel gain, enabling an offline solution approach in contrast to this work. Throughput maximization of WPT devices was previously considered in [47] where offline and online policies were presented in the context of a cognitive radio (CR) setting.

In this work, we investigate the problem of sensing optimization over a finite horizon in a WPCN where a WPD harvests energy from WPT of the AP tn sense a data packet at a specific resolution and then allocates the harvested energy in the subsequent time slots to transmit its data. Unlike the previous works, we consider a scenario where the CSI evolves randomly over the duration of the deadline, and CSI is only causally available at the transmitter which necessitates an online optimization framework. We avoid the complexity of the tabular methods (such as value iteration algorithm [36]) by deriving closed form solutions for the optimal WPT duration and power allocations in the IT period. We show how the simple closed-form expressions simplify addressing the sensing optimization problem. We address the sensing optimization problem in a reinforcement learning framework, where the optimum sensing resolution is learned by the WPD in a sequence of actions and observations. Finally, we conduct extensive simulations to verify our analytical findings.

D. OUTLINE

The paper is organized as follows: In Section II, we formally present the system model and all relevant assumptions. In Section III, we formulate the problem of sensing optimization. In Section IV, we formulate the sub-problem of finite horizon throughput maximization. In Section IV-A, we provide an upper bound on the maximum achievable throughput by assuming non-causal information. In Section IV-B, we solve the online counter-part of the problem by assuming only causal information. In Section V, we address the sensing optimization problem and in Section VI, we provide

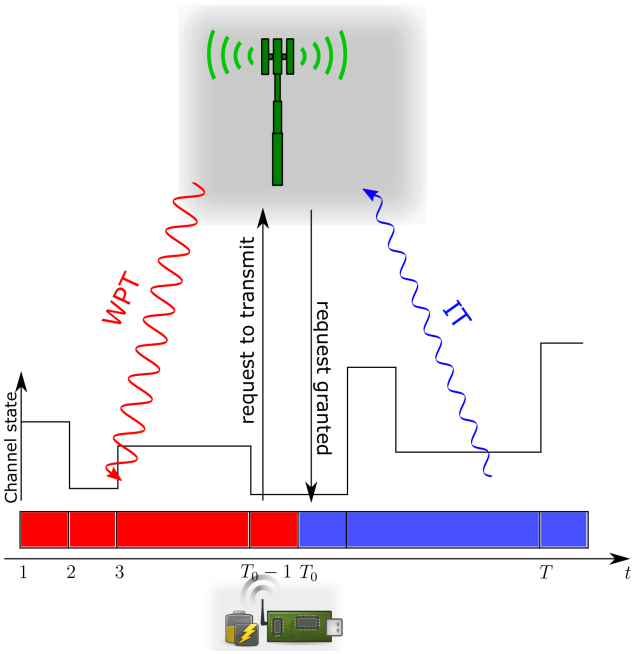


FIGURE 1. System model.

Monte-Carlo simulations to verify our findings. Finally, we conclude the paper in Section VII.

II. SYSTEM MODEL

We consider a point-to-point communication wireless channel where a WPD sends its sensory data to an AP by dynamically allocating power as shown in Figure 1. The AP uses WPT to replenish the battery of the WPD. The WPT and information transmission (IT) periods are non-overlapping in time, assuming a half-duplex transmission scenario. We consider a harvest-first-transmit-later policy where the WPD harvests energy for a certain duration and utilizes it to sense and transmit data to the AP. Such a policy eliminates the need for signaling between the sensor and the AP at each time slot and, hence, is more suitable for energy deprived sensors. The sensory unit of the WPD is capable of capturing data at K distinct resolution settings, each representing a quality point which is described by the number of bits used. Let L_k be the size of the type $k = 1, \ldots, K$ sensed data in bits. The duration of WPT and IT periods is governed by the channel gain process which jointly affects the amounts of the harvested energy and transmitted data. We assume a discrete time scenario over a finite horizon. The time is slotted $t = 1, \ldots, T$ and $T < \infty$ denotes the frame length in units of slots. Let $g(t)$, $E_h(t)$ be the channel gain, and the amount of harvested energy at time slot t, respectively. Specifically, the amount of harvested energy at time slot t is available at the beginning of slot $t + 1$. The wireless channel is modeled as a multi state independent and identically distributed (iid) random process with N levels. The channel gain remains constant for a duration of a time slot but changes randomly from one time slot to another, e.g., a wearable sensor exposed to blockage due to the movement of a person.

Modulation	Bits/symbol	SNR/symbol
2 PAM	1	$0.25d^2$
4 QAM	2	$0.50d^2$
16 QAM	4	$1.25d^2$
64 QAM	6	$5.25d^2$

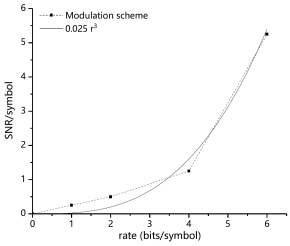


FIGURE 2. The comparison of monomial and actual transmission rate and required signal-to-noise (SNR) ratio per symbol for $m = 3$ and $\lambda = 0.025$ as given in [42]. d represents the minimum distance between signal points.

Let $g(t) \in \{g_1, \ldots, g_N\}$ be the channel power gain at slot t. We set $\mathbb{P}(g(t) = g_n) = q_n$.[1] The WPD only has causal CSI and only during the IT period.

The AP transmits a power beacon of P watts over the wireless channel for a duration of $T_0 - 1$ time slots. Assuming channel reciprocity, the amount of energy harvested by the WPD at time t is $E_h(t) = \eta\delta g(t)P$, where η is a constant representing the efficiency of the EH process[2] and δ is the duration of a time slot. The energy state of the WPD at time slot t is denoted by $E(t)$. Let us denote $e_n = \eta\delta g_n P$ as the amount of harvested energy when the channel state is at level n. At the beginning of the T_0-th time slot, the WPD consumes $\mathcal{E}_k$ Joules to sense L_k bits of data to be sent to the AP. Immediately after sensing the data, IT period starts.

At time slot $t \geq T_0$, the WPD transmits with power $p(t)$, and the received power at the AP is $p(t)g(t)$. In order to develop a tractable analytical solution, we assume a widely used empirical transmission energy model as in [37]–[42]. Specifically, the instantaneous rate of transmitting with power $p(t)$ when the channel gain is $g(t)$ is calculated by

$$r(t) = \sqrt[m]{\frac{p(t)g(t)}{\lambda}} \tag{1}$$

where λ denotes the energy coefficient incorporating the effects of bandwidth and noise power and m is the monomial order determined by the adopted coding scheme [42]. Figure 2 [42], compares the actual transmission rate with the monomial model described in (1). The approximated energy rate model, although may not be general for all cases, provides closed-form solutions for a challenging dynamic problem that gives insights to a practical and emerging problem.

Each type k data corresponds to a application specific utility upon being delivered to the AP. If the WPD successfully delivers a type k data, it receives a known utility of $Z(L_k)$, and zero otherwise. We emphasize that providing a high resolution input data provides a higher utility. However, the increased utility in the application layer comes at a price of reduced chance of delivering the input data to the AP due to the finite time horizon and the dynamic nature of the wireless link. Hence, there exists an optimal trade-off in balancing the quality of input data and probability of delivering it successfully to the AP for processing. The WPD aims at maximizing its utility by jointly determining the optimal sensing resolution; optimal WPT period duration, T_0; and optimal power allocation in the IT period, $p(t)$ for $t = T_0, \ldots, T$ in a decentralized fashion.

[1]Note that g_n's can be obtained by discretizing a continuous time channel process.

[2]Note that η in practice is a function of the received power and cannot be assumed to be a constant. We will show in Section IV how to extend the results to account for an η when it is a function of the received power.

III. PROBLEM FORMULATION

In this section, we formulate a joint utility optimization problem that aims at finding the optimal sensing resolution, the optimal trade-off between the EH and IT periods, and the dynamic control of transmission power during the IT period. More specifically, we aim at solving the following optimization problem.

$$\max_{L_k, T_0, \{p(t)\}_{t=T_0}^{T}} \quad Z(L_k)\mathbb{P}\left(\sum_{t=T_0}^{T} \sqrt[m]{\frac{g(t)p(t)}{\lambda}} > L_k\right) \tag{2}$$

$$p(t) \leq E(t)/\delta, \qquad t = T_0, \ldots, T, \tag{3}$$

$$E(t+1) = E(t) + E_h(t), \qquad t = 1, \ldots, T_0 - 1, \tag{4}$$

$$E(t+1) = E(t) - p(t)\delta - \mathcal{E}_k \mathbb{1}_{t=T_0}, \qquad t = T_0, \ldots, T, \tag{5}$$

$$L_{min} \leq L_k \leq L_{max}. \tag{6}$$

Note that (2) is the expected utility of delivering data type k, (3) ensures that the consumed energy does not exceed the available energy, (4) and (5) are the battery dynamics in the WPT and IT periods, and (6) is corresponds to the number of available resolution settings, respectively. Note that, in general, providing an explicit equation for $Z(L_k)$ may render infeasible as in the case of relating the error of estimating the gaze location to the number of pixels per frame. However, as we demonstrate in Section V, there is no need to have an explicit formulation for the utility function to optimize the sensing resolution. As long as there is a quantifiable mapping, either empirically or analytically, between L_k and the utility, we can find the optimal solution.

The above optimization problem consists of three subproblems; choosing the size of the input data L_k, determining the optimal WPT duration T_0, and optimal power allocations in the IT period $p(t)$, $t = T_0, \ldots, T$. Note that a policy which maximizes the expected throughput of the WPD, by optimizing the optimal WPT duration and power allocation in the IT period, has a better probability of success compared to any alternative policy. Thus, in the following, we first consider finite horizon throughput maximization by optimizing the WPT duration as well as power allocation in the IT period.

IV. FINITE HORIZON THROUGHPUT MAXIMIZATION

In this section, we jointly optimize the WPT duration and power allocation in order to maximize the expected throughput of the WPD. Explicitly, We aim at solving the following

optimization problem:[3]

$$\max_{T_0,\{p(t)\}_{t=T_0}^{T}} \quad \sum_{t=T_0}^{T} \sqrt[m]{\frac{g(t)p(t)}{\lambda}} \tag{7}$$

$$p(t) \le E(t)/\delta, \quad t = T_0, \dots, T, \tag{8}$$

$$E(t+1) = E(t) + E_h(t), \quad t = 1, \dots, T_0 - 1 \tag{9}$$

$$E(t+1) = E(t) - p(t)\delta, \quad t = T_0, \dots, T. \tag{10}$$

Note that the objective function (7) is the total number of transmitted bits in the IT period, (8) ensures that the consumed energy does not exceed the available energy, (9) and (10) are the battery dynamics in the WPT and IT periods, respectively. We first solve the offline version of the optimization problem by assuming that the channel gains are available prior to the optimization. Using the insights from the offline problem, we will design an optimal online policy, where the channel gains are only available causally.

A. OPTIMAL OFFLINE POLICY

We consider the offline counterpart of the optimization problem in (7). Thus, we assume that values of $g(t)$ are known non-causally for $t = 1, \dots, T$. Assuming that the optimal value of T_0 is given, we first aim at optimizing the power allocation in the IT period. We are interested in maximizing the following function

$$\max_{p(t)} \quad \sum_{t=T_0}^{T} r(t)$$
$$0 \le p(t) \le E(t)/\delta.$$

In Theorem 1, we show that the optimal policy, that maximizes the total number of bits transmitted bits in the IT period, allocates at each time slot a fraction of the available energy which depends on the current channel realization as well as a measure of future channel expectations.

Theorem 1: For a given T_0 and realizations of $g(t)$ for $t = 1, \dots, T$, the optimal dynamic power allocation for the offline problem is calculated by

$$p^*(t) = \frac{g(t)^{\frac{1}{m-1}}}{g(t)^{\frac{1}{m-1}} + G(t+1)^{\frac{1}{m-1}}} \frac{E(t)}{\delta} \tag{11}$$

where

$$G(t) = \begin{cases} \left[g(t)^{\frac{1}{m-1}} + G(t+1)^{\frac{1}{m-1}} \right]^{m-1}, & \text{if } t \le T \\ 0, & \text{if } t > T \end{cases}, \tag{12}$$

and the maximum number of transmitted bits is calculated as

$$\sum_{t=T_0}^{T} r^*(t) = \sqrt[m]{\frac{E(T_0)}{\delta\lambda} G(T_0)} \tag{13}$$

Proof: The proof follows DP backward recursion [44]. The proof is given in Appendix A. □

The offline optimization problem becomes:

$$\max_{T_0} \quad \sqrt[m]{\frac{E(T_0)}{\delta\lambda} G(T_0)} \tag{14}$$
$$2 \le T_0 \le T.$$

[3]For clarity of the presentation, we neglect the energy consumption of sensing, i.e, $\mathcal{E}_k$s, without affecting the main results. We consider them in the numerical evaluations.

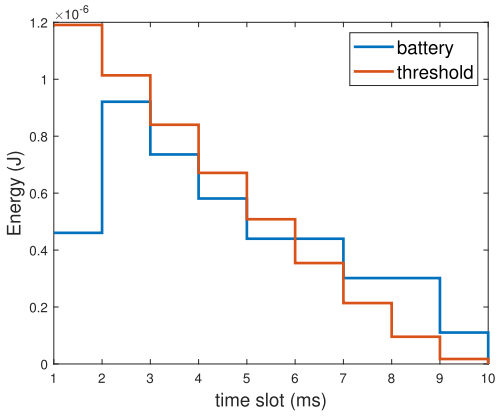


FIGURE 3. An illustrative example of the battery evolution, $E(t)$, where $T = 10$.

The above maximization problem has only one integer variable and hence, the optimal value for T_0 can be easily calculated numerically. In Figure 3, we illustrate a sample realization of the battery of the WPD. The time frame has 10 time slots, each with a duration of $1ms$. The WPD accumulates energy until $t = 2$. At $t = 3$, since the available energy is larger than the threshold, the WPT period is stopped and the IT period began.[4]

B. OPTIMAL ONLINE POLICY

Note that, in the online case, $g(t)$ is only available causally. Therefore, the optimization problem in (7)-(10) cannot be solved using offline optimization tools and an online algorithm is required for its solution. A common approach to solve similar problems is to use dynamic programming (DP) to find the solution numerically, and store the optimal decisions in a look-up table for the WPD. However, solving a DP and storing the result is prohibitive for resource constrained WPDs. In the following, we extend the insights gained in the offline case to the online counterpart of the optimization problem in (7).

At each time slot $t \ge T_0$, the WPD allocates a fraction of its remaining energy and allocates $p(t) = \alpha(t)E(t)/\delta$ as its transmit power. Hence, the optimization problem converts to:

$$\max_{T_0,\{\alpha(t)\}_{t=T_0}^{T}} \quad \sum_{t=T_0}^{T} \sqrt[m]{\frac{g(t)\alpha(t)E(t)}{\delta\lambda}} \tag{15}$$

$$0 \le \alpha(t) \le 1, \quad t = T_0, \dots, T, \tag{16}$$

$$E(t+1) = E(t) + E_h(t), \quad t = 1, \dots, T_0 - 1 \tag{17}$$

$$E(t+1) = (1 - \alpha(t))E(t), \quad t = T_0, \dots, T. \tag{18}$$

1) DYNAMIC ENERGY ALLOCATION

In this section, we first optimize the values of $\alpha(t)$ by conditioning on T_0. Then using the obtained result, we will give a criteria for stopping the EH process, i.e., optimizing the value of T_0.

[4]In Section IV-B, we show how to calculate the optimal WPT duration and power allocations in the IT period.

Let the IT period begin at T_0 and aim to maximize the throughput over $T - T_0$ time slots by using DP. The problem is recursively solved starting at the last time slot T, and the result is propagated by recursion until it reaches $t = T_0$. We denote the instantaneous reward of choosing $\alpha(t)$ by $U_{\alpha(t)}(E(t), g(t))$ which is the instantaneous number of bits transmitted to the AP, when the the amount of available energy at time t, is $E(t)$ and the channel power gain is at state $g(t)$. Thus,

$$U_{\alpha(t)}(E(t), g(t)) = \sqrt[m]{\frac{\alpha(t)g(t)E(t)}{\delta\lambda}}. \tag{19}$$

We denote the action-value function by $V_\alpha(E(t), g(t))$ which is equal to the instantaneous reward of choosing $\alpha(t)$ plus the expected number of bits that can be transmitted in the future. Hence, the action-value function evolves as,

$$V_{\alpha(t)}(E(t),g(t)) = U_{\alpha(t)}(E(t), g(t)) + \sum_{i=1}^{N} q_i V(E(t+1),g_i), \tag{20}$$

where, $V(E(t), g(t))$ is the value function defined as,

$$V(E(t), g(t)) = \max_{\alpha(t)} V_{\alpha(t)}(E(t), g(t)). \tag{21}$$

Note that at the last time slot, i.e., $t = T$, all the energy in the battery will be used for transmission, i.e., $\alpha(T) = 1$. Thus, it follows that,

$$\begin{aligned} V(E(T), g(t)) &= U_1(E(T), g(T)) \\ &= \sqrt[m]{\frac{g(T)E(T)}{\delta\lambda}} \\ &= \sqrt[m]{\frac{g(T)(1-\alpha(T-1))E(T-1)}{\delta\lambda}}. \end{aligned} \tag{22}$$

We maximize the action-value function at $t = T - 1$ by optimizing $\alpha(T-1)$ as follows,

$$\begin{aligned} &V_\alpha(E(T-1), g(T-1)) \\ &\quad = U_\alpha(E(T-1), g(T-1)) \\ &\qquad + \sum_{i=1}^{N} q_i V((1-\alpha(T-1))E(T-1), g_i) \\ &\quad = \sqrt[m]{\frac{g(T-1)\alpha(T-1)E(T-1)}{\delta\lambda}} \\ &\qquad + \sum_{i=1}^{N} q_i \sqrt[m]{\frac{g_i((1-\alpha(T-1))E(T-1))}{\delta\lambda}}. \end{aligned} \tag{23}$$

It is easy to see that (23) is concave with respect to $\alpha(T-1)$. Therefore, by differentiating (23), the optimal $\alpha(T-1)$ can be calculated as follows:

$$\alpha^*(T-1) = \frac{g(T-1)^{\frac{1}{m-1}}}{g(T-1)^{\frac{1}{m-1}} + Q(T-1)^{\frac{m}{m-1}}}, \tag{24}$$

where,

$$Q(T-1) = \sum_{i=1}^{N} q_i \sqrt[m]{g_i}. \tag{25}$$

The corresponding value function can also be calculated as

$$\begin{aligned} V(E(T-1), g(T-1)) = \sqrt[m]{\frac{E(T-1)}{\delta\lambda}} \big(g(T-1)^{\frac{1}{m-1}} \\ + Q(T-1)^{\frac{m}{m-1}}\big)^{\frac{m-1}{m}}. \end{aligned} \tag{26}$$

In a similar manner as above, we can recursively calculate the optimal $\alpha(t)$ for $t = T-2, \ldots, T_0$. The result is summarized in the following theorem.

Theorem 2: For any $t = T-1, \ldots, T_0$, the optimal decision is to choose

$$\alpha^*(t) = \frac{g(t)^{\frac{1}{m-1}}}{g(t)^{\frac{1}{m-1}} + Q(t)^{\frac{m}{m-1}}}, \tag{27}$$

where

$$Q(t) = \sum_{i=1}^{N} q_i \big(g_i^{\frac{1}{m-1}} + Q(t+1)^{\frac{m}{m-1}}\big)^{\frac{m-1}{m}}. \tag{28}$$

The corresponding value function is

$$V(E(t), g(t)) = \sqrt[m]{\frac{E(t)}{\delta\lambda}} \big(g(t)^{\frac{1}{m-1}} + Q(t)^{\frac{m}{m-1}}\big)^{\frac{m-1}{m}} \tag{29}$$

Proof: The proof is given in Appendix B. □

Theorem 2 gives a framework to dynamically allocate energy at each time slot $t \geq T_0$. Instead of numerically solving the DP and storing it in a large look up table, WPD needs to just calculate and store an array of values with a maximum dimension of T. The closed form expressions derived in (27)-(29) significantly simplify the procedure to optimize T_0. We will use these results to find an structure for the optimal stopping time problem in the subsequent section.

2) OPTIMAL STOPPING TIME FOR THE WPT DURATION

In the following, we derive the optimal stopping time for the WPT duration, i.e., optimizing T_0 in (7)-(10). Recall that the WPD accumulates energy up to some time t, and then stops the WPT to start transmitting its data bits. Also, recall that during WPT, the WPD is blind to the channel conditions. If the WPD stops the WPT at time t, then the expected number of bits that can be transmitted is

$$\begin{aligned} \sum_{i=1}^{N} q_i V(E(t), g_i) &= \sum_{i=1}^{N} q_i \sqrt[m]{\frac{E(t)}{\delta\lambda}} \big(g_i^{\frac{1}{m-1}} + Q(t)^{\frac{m}{m-1}}\big)^{\frac{m-1}{m}} \\ &= \sqrt[m]{\frac{E(t)}{\delta\lambda}} Q(t-1). \end{aligned} \tag{30}$$

Note that (30) follows from the definition of $Q(t)$ given in (28).

Let $J_t(E(t))$, $t = 1, \ldots, T$ be the maximum expected number of bits that can be transmitted if the WPT is stopped at time t, and the amount of available energy is $E(t)$. At any time t, the WPD will either stop or continue the WPT. The optimal stopping time for the WPT can be formulated as

$$\max_{t \leq T} J_t(E(t)), \tag{31}$$

where

$$J_t(E(t)) = \max\left(\sqrt[m]{\frac{E(t)}{\delta\lambda}}Q(t-1), \mathbb{E}(J_{t+1}(E(t+1))\Big|E(t))\right). \quad (32)$$

The problem can be formulated as a DP and recursively solved for every possible $E(t)$ and t. Before proceeding, we need the following lemma.

Lemma 1: $Q(t)$, defined in (28) is a monotonically decreasing function in t.

Proof:

$$\frac{Q(t)}{Q(t+1)} = \frac{\sum_{i=1}^{N} q_i\left(g_i^{\frac{1}{m-1}} + Q(t+1)^{\frac{m}{m-1}}\right)^{\frac{m-1}{m}}}{Q(t+1)} = \sum_{i=1}^{N} q_i\left(1 + \frac{g_i^{\frac{1}{m-1}}}{Q(t+1)^{\frac{m}{m-1}}}\right)^{\frac{m-1}{m}} > 1. \quad (33)$$

It readily follows that $Q(t) > Q(t+1)$. □

Note that at $t = T$, the best strategy is to stop the WPT and start the IT period, since otherwise no bits can be transmitted to the AP. Thus,

$$J_T(E(T)) = \sqrt[m]{\frac{E(T)}{\delta\lambda}}Q(T-1). \quad (34)$$

We continue the recursive evaluation at time slot $t = T-1$. We have,

$$\begin{aligned} &J_{T-1}(E(T-1)) \\ &= \max(\sqrt[m]{\frac{E(T-1)}{\delta\lambda}}Q(T-2), \mathbb{E}(J_T(E(T))|E(T-1))) \\ &= \max(\sqrt[m]{\frac{E(T-1)}{\delta\lambda}}Q(T-2), \\ &\quad \sum_{i=1}^{N} q_i \sqrt[m]{\frac{E(T-1)+e_i}{\delta\lambda}}Q(T-1)) \end{aligned} \quad (35)$$

Since $Q(T-2) > Q(T-1)$ as proven in Lemma 1, if $E(T-1) \geq \gamma(T-1)$, then

$$\sqrt[m]{\frac{E(T-1)}{\delta\lambda}}Q(T-2) \geq \sum_{i=1}^{N} q_i \sqrt[m]{\frac{E(T-1)+e_i}{\delta\lambda}}Q(T-1)), \quad (36)$$

where $\gamma(T-1)$ is the solution to the following equation

$$\sum_{i=1}^{N} q_i \sqrt[m]{1 + \frac{e_i}{\gamma(T-1)}} = \frac{Q(T-2)}{Q(T-1)}. \quad (37)$$

Note that $\gamma(T-1)$ admits a unique solution because the left hand side of (37) is a strictly decreasing function in $\gamma(T-1)$ and its range belongs to $(1, \infty)$. Also, from Lemma 1, we know that $\frac{Q(T-2)}{Q(T-1)} > 1$. Hence, it is optimal to stop the WPT at time $T-1$ if $E(T-1) \geq \gamma(T-1)$. This suggests that the optimal stopping times are governed by a time varying threshold type structure, where at any given time t, it is optimal to stop the WPT if $E(t) \geq \gamma(t)$. Before, proving this observation, we need the following lemma.

Lemma 2: For any $k = 1, \ldots, T-1$, we have

$$\frac{Q(T-k-1)}{Q(T-k)} < \frac{Q(T-k)}{Q(T-k+1)} \quad (38)$$

Proof: By using (28), we have

$$\frac{Q(T-k-1)}{Q(T-k)} = \frac{\sum_{i=1}^{N} q_i\left(g_i^{\frac{1}{m-1}} + Q(T-k)^{\frac{m}{m-1}}\right)^{\frac{m-1}{m}}}{Q(T-k)} = \sum_{i=1}^{N} q_i\left(1 + \frac{g_i^{\frac{1}{m-1}}}{Q(T-k)^{\frac{m}{m-1}}}\right)^{\frac{m-1}{m}}, \quad (39)$$

and,

$$\frac{Q(T-k)}{Q(T-k+1)} = \frac{\sum_{i=1}^{N} q_i\left(g_i^{\frac{1}{m-1}} + Q(T-k+1)^{\frac{m}{m-1}}\right)^{\frac{m-1}{m}}}{Q(T-k+1)} = \sum_{i=1}^{N} q_i\left(1 + \frac{g_i^{\frac{1}{m-1}}}{Q(T-k+1)^{\frac{m}{m-1}}}\right)^{\frac{m-1}{m}}. \quad (40)$$

From Lemma 1, we have $Q(T-k) > Q(T-k+1)$ and thus the lemma holds. □

In the following theorem, we give the structure of the optimal stopping policy.

Theorem 3: At each time slot t, the optimal decision is to stop the WPT if $E(t) \geq \gamma(t)$, where $\gamma(t)$ is the solution to the following equation,

$$\sum_{n=1}^{N} q_n \sqrt[m]{1 + \frac{e_n}{\gamma(t)}} = \frac{Q(t-1)}{Q(t)} \quad (41)$$

Proof: The proof is in Appendix C. □

Note that the results of Theorem 3 can be easily extended to account for the dependability of EH efficiency, η, on the received power. More specifically, when the amount of harvested energy at fading state n is defined to be $e_n = \eta(g_nP)g_nP$, where $\eta(g_nP)$ is the EH efficiency when the received power at the WPD is g_nP, all the derivations given in the paper remain valid.

The results established in Theorem 2 and 3 enables us to develop an online low complexity optimal algorithm that maximizes the expected throughput. The procedure is summarized in Algorithm 1.

Algorithm 1 Online Policy

1: Initialize $Q(t)$ for $t = 0, \ldots, T-1$ using (28),
2: Initialize $\gamma(t)$ for $t = 1, \ldots, T-1$ using (41),
3: **for** $t = 1 : T$ **do**
4: **if** $E(t) < \gamma(t)$ **then**
5: continue the WPT
6: **else**
7: $T_0 = t$,
8: Stop the WPT,
9: Break
10: **for** $t = T_0 : T$ **do**
11: Calculate $\alpha(t)$ using (27),
12: Transmit using $\alpha(t)E(t)$.

The time complexity of Algorithm 1 only depends on line 1 and 2 and the rest of the algorithm has a constant time

complexity with respect to N, and T. Line 1 solves (28) where a constant time operation (i.e., the term inside the summation) is evaluated N times for any given $t = 1, \ldots, T-1$. Since (28) is evaluated T times, the complexity of line 1 is at most $O(NT)$. Line 2 calculates the thresholds by solving (41). Consider a root finding algorithm which solves (41) by evaluating the function at different points (e.g., bisection method). Since (41) involves summation of N nonlinear functions, the root finding algorithm needs to evaluate values of N non-linear functions. Thus, for a given t the complexity is $O(N)$. Moreover, since it is calculated at most T times, the overall complexity is $O(NT)$. Thus the overall complexity of Algorithm 1 is $O(NT)$. It is worth mentioning that if the statics of the channel do not change over time, line 1 and 2 need to be calculated only once.

Remark 1: Note that that the monomial rate function have enabled a closed form solution of the optimal power allocations and WPT duration. However, it is also possible to extend this work beyond the monomial rate function to the Shannon rate function. The optimal solutions for power allocations and WPT duration can be derived with the same recursive approach presented in this section. However, due to the logarithmic nature of the Shannon rate function, it is no longer possible to derive closed form solutions, and thus, we have to resort for tabular methods to store the optimal solutions. For each possible state, $(E(t), g(t), t)$, the optimal power allocation $p(t)$ should be calculated and stored in the table. A similar table is also required for storing the optimal duration of WPT. An obvious drawback of the tabular method is that the WPD endures significant computational complexities as well as memory requirement due to the large number of states.

V. OPTIMAL SENSING

Thus far, we have developed a policy that maximizes the expected finite horizon throughput of the WPD by determining the optimal WPT duration and dynamic power allocation in a distributed manner. Recall that the ultimate goal is to maximize the sensing utility of the WPD by optimizing the sensing resolution. Algorithm 1, is a framework that maximizes the chance of successful delivery of the data to the AP. Thus the last quantity to be optimized is the sensing resolution.

Remark 2: Note that it is possible to increase the efficiency of the sensing utility by further compressing the sensed packets prior to the transmission as in [46]. Compressing the sensed packets decreases the number bits per packet and thus, increases the chance of delivering the packet. At the same time, due to utilizing the CPU for a number of cycles, the energy consumption of the WPD increases because of the compression. Hence, there exists a trade-off between the size of the sensed packet, compression ratio and the extra energy consumption. Compression can be easily accounted for in the learning model by simply extending the action space of the WPD to account for the compression ratio. We note that the inherent trade-off in compression is similar to the sensing resolution, and it can be incorporated in the formulation in a straightforward manner.

Let the event of successfully delivering a packet of L_k bits be χ_k. More specifically:

$$\chi_k = \begin{cases} 1 & \text{if } \sum_{t=T_0}^{T} r(t) > L_k, \\ 0 & \text{otherwise.} \end{cases} \tag{42}$$

We rewrite the optimization problem of interest as follows[5]

$$\max_{\{L_k\}_{k=1}^{K}} Z(L_k)\mathbb{E}(\chi_k) \tag{43}$$

The WPD in the beginning of each transmission frame chooses a L_k that optimizes the above optimization problem.[6] The unknown quantities in the optimization problem are $\mathbb{E}(\chi_k)$, $k = 1, \ldots, K$. We aim to learn these quantities using a reinforcement learning (RL) technique. The RL framework interacts with the environment and learns the values of the parameters of interest by observing the outcomes of its decisions. Note that the observation feedbacks are limited and only the feedback associated with the chosen decision in a time slot is observed. This problem can be efficiently formulated in the context of multi armed bandit (MAB) problem. The parameters of interest in the MAB are denoted by $\theta_k = \mathbb{E}(\chi_k) = \mathbb{P}(\chi_k = 1)$. We aim to efficiently infer each θ_k by interacting with the environment and observing the outcomes. In a MAB there are multiple arms (i.e., actions) each generating a random reward according to a probability distribution function (PDF). An agent sequentially chooses an action $x_t = k$ for $t = 1, \ldots$ and readjusts it strategy by observing the reward with the hope of maximizing its expected reward. In our problem, there are K actions. The WPD keeps initial estimates of $\hat{\theta}_k$ about the unknown parameters θ_k. The WPD chooses an action $x_t = k$ and observes the event $Z(L_k) \cdot \chi_k$. Based on the observation, it updates $\hat{\theta}_k$ until the algorithm converges to the optimal value. The typical method for optimizing a MAB problem is by the well known ϵ-greedy algorithm presented in Algorithm 2. The ϵ-greedy algorithm consists of two steps; exploration and exploitation. Exploration improves the estimate of non-greedy actions' values while exploitation is favorable when we reach a sufficient knowledge about the estimate of actions. ϵ-greedy algorithm, with probability (w.p.) $1-\epsilon$, greedily chooses an action k that maximizes $Z(L_k)\hat{\theta}_k$ and w.p. ϵ randomly chooses an action. In other words, w.p. ϵ the algorithm explores in the action space of the MAB while w.p. $1-\epsilon$ the algorithm exploits what it already knows. Although such an approach is guaranteed to approach the optimal performance [36], provided that ϵ is sufficiently small, the convergence rate of the algorithm

[5]The sensing formulation can be generalized beyond the indicator function for a utility function generating rewards with a support in [0, 1].

[6]Note that a better strategy is to choose the size of the data after observing the amount of harvested energy and the duration of IT period. Since the amount of harvested energy is independent upon each observation, we can easily extend the framework by considering a contextual multi armed bandit problem.

is poor. This is because ϵ-greedy algorithm does not judiciously explore in the parameter space. To speed up the convergence, we use a Bayesian inference method to judiciously explore in the action space of the MAB problem. The augmentation of the Bayesian framework in MAB is known as Thompson sampling (TS)[7] [43]. To see how TS works, let us model the uncertainty θ_k by assuming a prior distribution for it. Each θ_k is distributed according to a Beta distribution with parameters a_k and b_k. In particular, for each arm k, the prior probability density function of θ_k is:

$$\mathbb{P}(\theta_k) = \frac{\Gamma(a_k + b_k)}{\Gamma(a_k)\Gamma(b_k)} \theta_k^{a_k - 1}(1-\theta_k)^{b_k - 1}, \tag{44}$$

where $\Gamma(.)$ denotes the gamma function. The reason for choosing Beta as prior distribution is the conjugacy property of Beta distribution with Bernoulli distribution. In other words, if prior is Beta distributed and the likelihood is Bernoulli distributed, then the posterior distribution is also Beta distributed. This facilitates the process of sampling from the posterior distribution[8] Given a sample realization of χ_k, we are interested in updating the posterior distribution of θ_k. We have:

$$\begin{aligned}\mathbb{P}(\theta_k|\chi_k) &\propto \mathbb{P}(\theta_k)\mathbb{P}(\chi_k|\theta_k) \\ &= \frac{\theta_k^{a_k-1}(1-\theta_k)^{b_k-1}}{B(a_k,b_k)} \theta_k^{\chi_k}(1-\theta_k)^{1-\chi_k} \\ &\propto \theta_k^{a_k-1+\chi_k}(1-\theta_k)^{b_k-1+1-\chi_k}\end{aligned} \tag{45}$$

Hence, the posterior distribution is also Beta distributed with parameters, $a_k + \mathbb{1}_{\{\chi_k=1\}}$ and $b_k + \mathbb{1}_{\{\chi_k=0\}}$. Note that at any given time, only a single observation regarding the chosen action is revealed. Hence, after retrieving the observation about an action, the parameters of the posterior distribution is updated as:

$$(a_k, b_k) \leftarrow \begin{cases} (a_k, b_k) & \text{if } x_t \neq k, \\ (a_k + \chi_k, b_k + 1 - \chi_k) & \text{if } x_t = k. \end{cases} \tag{46}$$

The TS algorithm is given in Algorithm 3. Note that the only difference between the TS and ϵ-greedy algorithms in the exploration phase of the problem. TS judiciously explores by modeling the uncertainty of each action using a distribution with decreasing variance in the number of observations explored. This prevents the TS from exploring the actions that are believed to be sub-optimal. Meanwhile ϵ-greedy explores the action space randomly, reducing the efficiency of the exploration phase.

VI. NUMERICAL RESULTS

In this section, we compare the performance of the optimal online policy with that of the offline as well as two benchmark policies, namely uniform and power-halving policies.

[7]See [48]–[50] for the optimality analysis of TS.

[8]Note that the conjugacy property only makes it easier to sample from the posterior distribution. In case where the posterior distribution does not admit any known PDF, efficient Monte-Carlo methods such as Markov chain Monte-Carlo (MCMC) [45] method and its variants such as Gibbs sampling can be used to efficiently sample from the posterior.

Algorithm 2 ϵ-Greedy

for $t = 1, 2, \ldots$ **do**
  With probability ϵ
  for k=1,...,K **do**
    $\hat{\theta}_k = \frac{a_k}{a_k + b_k}$
  $x_t \leftarrow \begin{cases} \arg\max_k Z(L_k)\hat{\theta}_k & \text{with prob. } 1-\epsilon, \\ \text{choose a random action} & \text{with prob. } \epsilon. \end{cases}$
  Apply x_t and observe χ_k
  update the posterior using (46)

Algorithm 3 Thompson Sampling (TS)

for $t = 1, 2, \ldots$ **do**
  Sample from the posterior
  for k=1,...,K **do**
    Sample $\hat{\theta}_k \sim beta(a_k, b_k)$
  $x_t \leftarrow \arg\max_k Z(L_k)\hat{\theta}_k$
  Apply x_t and observe χ_k
  update the posterior using (46)

In uniform policy, the amount of harvested energy is uniformly distributed in the IT period. Power-halving policy allocates half of its available energy in each time slot in the IT period. The WPT duration for both uniform and power-halving policies is optimized using exhaustive search method. We also evaluate the performance of TS algorithm in the sensing utility maximization problem developed in Section V and compare it with that of ϵ-greedy.

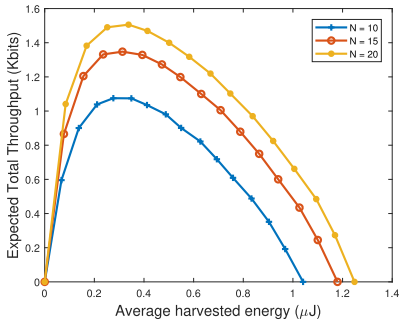
For the channel states, we assume a Rayleigh fading model with an average channel gain of 1. We assume that the AP transmits with power $P = 20$dBm which is normalized with respect to distance and EH efficiency. Time slot duration is 1ms, the bandwidth is assumed to be 2KHz, and the noise power density is 176 dBm/Hz.

A. RATE-ENERGY TRADE-OFF

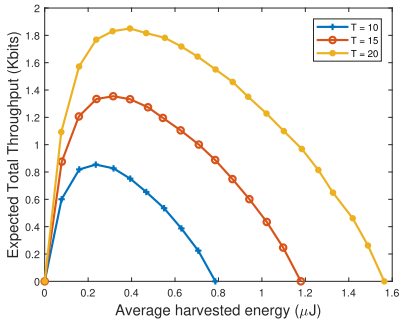
We first evaluate the rate-energy trade-off of the online policy which is the expected total number of bits transmitted with respect to the amount of harvested energy in a finite duration of T. In Figure 4a, for different values of channel discretization level, N, and a frame length of 15 time slots, the rate-energy trade-off is depicted. For different values of T, and $N = 15$, Figure 4b, illustrates the rate-energy trade-off. We observe from the figures that, spending too much time for transmitting more energy in the EH period reduces the time for IT period which in turn reduces the throughput. On the other hand, if we reduce the EH period, there would be less energy in the IT period resulting in a reduced throughput. Hence, an optimal balance is required.

B. PERFORMANCE EVALUATION

In Figure 5, when the fading is Rayleigh, the expected total number of bits that are transmitted in 100 time slots is depicted with respect to the number of channel discretization levels, N. We observe that as the number of channel levels



(a) Expected throughput with respect to N.



(b) Expected throughput with respect to T.

FIGURE 4. The effect of channel discretization and deadline duration on the expected throughput.

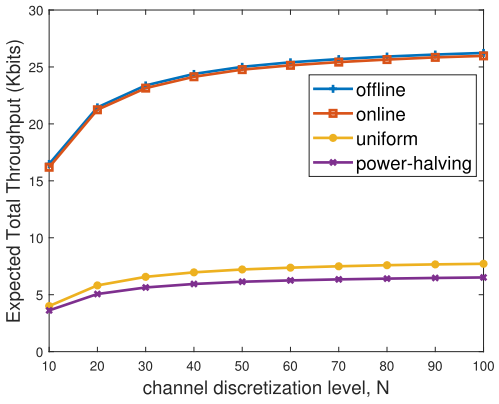


FIGURE 5. Expected total throughput of the WPD with respect to the number of channel discretization levels in $T = 100$ time slots.

increases, the discretization error decreases and hence the throughput of the all policies improve. The online policy achieves a throughput close to the upper-bound by optimally determining the WPT duration and power allocation in the IT period. Although the uniform and power-halving policies harvest energy for an optimum duration, they considerably perform poor due to the blind power allocation in the IT period.

Next, we plot the expected total throughput of the WPD in Figure 6. Again, the online policy, for all values of T, achieves an outstanding performance compared to the offline policies. For smaller values of T, the power-halving policy achieves a good performance. However, as T increases, due

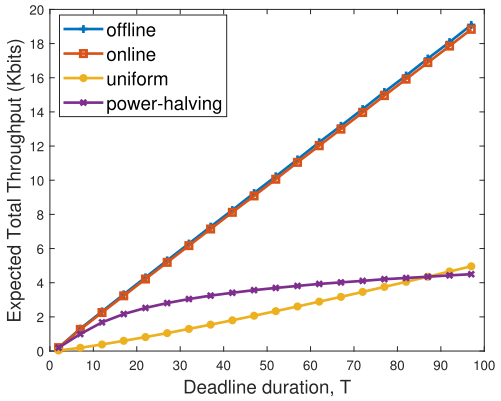


FIGURE 6. Expected total throughput of the WPD with $N = 20$ channel levels with respect to the frame length, T.

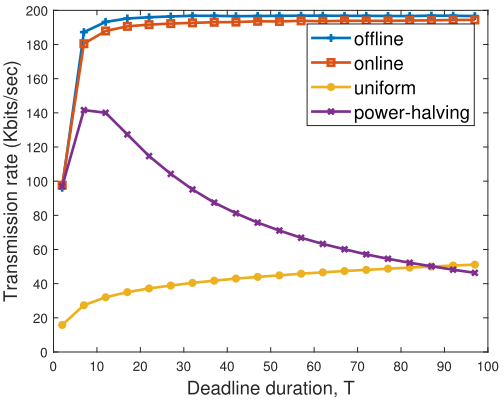


FIGURE 7. Expected transmission rate of the WPD with $N = 20$ channel levels with respect to the frame length, T.

to the concave nature of the rate-power function, the power-halving strategy becomes significantly inefficient. On the other hand, uniform policy is able to perform better, for larger values of T, with respect to power-halving policy by allocating the harvested energy uniformly across the IT period.

Finally, we illustrate the transmission rate of the WPD in units of bits per seconds (bits/sec) in Figure 7. IT can be seen that the online policy has a significantly higher rate than the uniform and power-halving policies. It is also evident that on the average, the online policy achieves a significantly good performance with respect to the offline policy.

C. MAB

Here, we evaluate the performance of TS and ϵ-greedy algorithms and compare their performance. In Figure 8, we plot the per-period regret of both algorithms. For plots, we use the following synthetic parameters; $T = 15$, $N = 30$, $L = 1000, 2500, 3000$ bits, $Z = 500, 700, 750$, and $\mathcal{E} = 1, 3, 4$ μJoules. Per-period regret is the gap between the optimal utility and the utility achieved by the given algorithm. We obtain the value of the optimal utility by exhaustive search for comparison purposes only. Each point in Figure 8 is averaged over 10^5 samples.

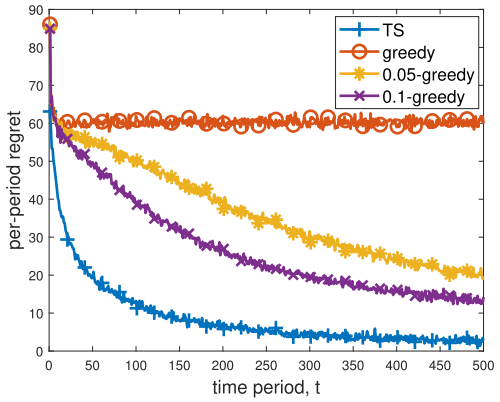


FIGURE 8. Per-period regret comparison of TS and ϵ-greedy algorithms for $\epsilon = 0, 0.05, 0.1$.

The greedy algorithm ($\epsilon = 0$) has the worst performance as it does not explore at all. By giving non-zero values for ϵ, we can see that 0.05-greedy and 0.1-greedy greatly improve upon the greedy algorithm by performing explorations. However, we see a poor performance regarding their convergence rate. TS improves the convergence rate significantly by simply adding intelligence to the exploration phase. This makes the TS algorithm to approach a per-period regret of 0 considerably faster than the ϵ-greedy algorithm.

VII. CONCLUSION

In this work, we studied a sensing optimization for a WPD operating in a finite horizon. The WPD harvests energy from the RF signals transmitted by an access point to sense data and transmit it to the AP for achieving a utility that depends on the quality of the sensed data. The wireless channel varies randomly over the horizon and it is only available to the WPD causally and only in the IT period. The achieved utility by the WPD depends on both the quality of the sensed data and the chance of delivering it to the AP. Therefore, we first maximized the probability of successful data delivery by optimizing the WPT duration and dynamic power allocation. To gain insight to the dynamic throughput maximization problem, we first studied the offline problem where we assumed the channel realizations are known non-causally. We then studied the online counterpart of the problem by assuming that the channel realization are available only causally and in the IT period. We show that there exist a time-dependent threshold on the energy level of the WPD in which it is optimal to stop WPT and start the IT period. Then, we show that the optimal power allocation in the IT period follows a fractional structure in which the WPD at each time slot allocates a fraction of its energy that depends on the immediate channel state as well as a measure of future expectations. The numerical results show that the online policy achieves a performance significantly close to the upper-bound. By using the throughput maximization results, we then formulated a Bayesian inference reinforcement learning problem to finally address the sensing resolution optimization problem. We show that the Bayesian inference achieves a convergence rate that is much faster than that of the ϵ-greedy algorithm. In the future, we aim to characterize the performance gap due to the use of the monomial rate function and the Shannon rate along with the extension of the network to multiple WPDs.

APPENDIX A
PROOF OF THEOREM 1

Consider the following concave optimization of the throughput at time $T-1$ and $T+1$, given that the amount of available energy at time $T-1$ is $E(T-1)$

$$\begin{aligned}\max_{p(T-1),p(T)} \quad & \sqrt[m]{\frac{g(T-1)p(T-1)}{\lambda}} \\ & + \sqrt[m]{\frac{g(T)(E(T-1)/\delta - p(T-1))}{\lambda}} \\ & p(T-1) \le E(T-1)/\delta, \\ & p(T) \le E(T-1)/\delta - p(T-1).\end{aligned}$$

The WPD at the last time slot should utilize all the available energy before the transmission frame expires. Hence, we set $p(T) = E(T-1)/\delta - p(T-1)$. The optimization problem becomes

$$\begin{aligned}\max_{p(T-1)} \quad & \sqrt[m]{\frac{g(T-1)p(T-1)}{\lambda}} + \sqrt[m]{\frac{g(T)(E(T-1)/\delta - p(T-1))}{\lambda}} \\ & 0 \le p(T-1) \le E(T-1)/\delta.\end{aligned}$$

The Lagrangian of the above problem can be written as

$$\begin{aligned}\mathcal{L}(p(T-1), \mu_1, \mu_2) = & \sqrt[m]{\frac{g(T-1)p(T-1)}{\lambda}} \\ & + \sqrt[m]{\frac{g(T)(E(T-1)/\delta - p(T-1))}{\lambda}} \\ & - \mu_1(p(T-1) - E(T-1)/\delta \\ & + \mu_2 p(T-1)\end{aligned}$$

The derivative of the Lagrangian is calculated as follows

$$\begin{aligned}\frac{\partial \mathcal{L}(p(T-1), \mu_1, \mu_2)}{\partial p(T-1)} = & \frac{1}{m}\sqrt[m]{\frac{g(T-1)}{\lambda}} p(T-1)^{\frac{1}{m}-1} \\ & - \frac{1}{m}\sqrt[m]{\frac{g(T)}{\lambda}} (E(T-1)/\delta \\ & - p(T-1))^{\frac{1}{m}-1} + (\mu_2 - \mu_1)\end{aligned}$$

Prior to equating the Lagrangian to zero, we assume that the optimal power allocation satisfies the constraint, i.e., $0 \le p^*(T-1) \le E(T-1)/\delta$, and set $\mu_1 = \mu_2 = 0$. By solving the derivative of the relaxed Lagrangian, we get

$$p^*(T-1) = \frac{g(T-1)^{\frac{1}{m-1}}}{g(T-1)^{\frac{1}{m-1}} + g(T)^{\frac{1}{m-1}}} E(T-1)/\delta$$

Note that since $0 \le \frac{g(T-1)^{\frac{1}{m-1}}}{g(T-1)^{\frac{1}{m-1}} + g(T)^{\frac{1}{m-1}}} \le 1$, the constraint is satisfied. Let us calculate the optimum sum throughput at

time $T-1$ and T:

$$\begin{aligned} r(T-1)+r(T) &= \sqrt[m]{\frac{g(T-1)p^*(T-1)}{\lambda}} \\ &\quad + \sqrt[m]{\frac{g(T)(E(T-1)/\delta - p^*(T-1))}{\lambda}} \\ &= \sqrt[m]{\frac{E(T-1)/\delta}{\lambda}}\left[\sqrt[m]{\frac{g(T-1)g(T-1)^{\frac{1}{m-1}}}{g(T-1)^{\frac{1}{m-1}}+g(T)^{\frac{1}{m-1}}}}\right. \\ &\quad \left. + \sqrt[m]{\frac{g(T)g(T)^{\frac{1}{m-1}}}{g(T-1)^{\frac{1}{m-1}}+g(T)^{\frac{1}{m-1}}}}\right] \\ &= \sqrt[m]{\frac{E(T-1)/\delta}{\lambda}}\frac{g(T-1)^{\frac{1}{m-1}}+g(T)^{\frac{1}{m-1}}}{\sqrt[m]{g(T-1)^{\frac{1}{m-1}}+g(T)^{\frac{1}{m-1}}}} \\ &= \sqrt[m]{\frac{E(T-1)/\delta G(T-1)}{\lambda}}, \end{aligned}$$

where $G(T-1)=\left[g(T-1)^{\frac{1}{m-1}}+g(T)^{\frac{1}{m-1}}\right]^{m-1}$. To generalize the results, we use induction. Suppose that the above results are true for some time $t+1$. Next consider the optimization of sum throughput from time t to T:

$$\max_{p(t)} \sqrt[m]{\frac{g(t)p(t)}{\lambda}} + \sqrt[m]{\frac{(E(t)/\delta - p(t))G(T-1)}{\lambda}}$$

Similar to the above analysis, it follows that

$$\begin{aligned} p^*(t) &= \frac{g(t)^{\frac{1}{m-1}}}{g(t)^{\frac{1}{m-1}}+G(t+1)^{\frac{1}{m-1}}}E(t) \\ \sum_{\tau=t}^{T} r(\tau) &= \sqrt[m]{\frac{E(t)G(t)}{\lambda}}, \end{aligned}$$

where $G(t)=\left[g(t)^{\frac{1}{m-1}}+G(t+1)^{\frac{1}{m-1}}\right]^{m-1}$.

APPENDIX B
PROOF OF THEOREM 2

The proof is by induction. We have shown in (24), (25), and (26), that the case for $k=1$ is true. By assuming the the case for $k-1$ is true, let us calculate the case k. The value function is given as

$$\begin{aligned} V_\alpha(E(T-k),g(T-k)) &= U_\alpha(E(T-k),g(T-k)) \\ &\quad + \sum q_i V(E(T-(k-1)),g_i) \end{aligned} \quad (47)$$

Note that $E(T-(k-1))=(1-\alpha(T-k))E(T-k)$ and since the case is true for $k-1$, from (29), we have

$$\begin{aligned} V(E(T-(k-1)),g_i) &= \sqrt[m]{\frac{(1-\alpha(T-k))E(T-k)/\delta}{\lambda}}\left(g_i^{\frac{1}{m-1}}\right. \\ &\quad \left. + Q(T-k+1)^{\frac{m}{m-1}}\right)^{\frac{m-1}{m}} \end{aligned} \quad (48)$$

By substituting (48) in (47) we get

$$\begin{aligned} V_\alpha(E(T-k),g(T-k)) &= \sqrt[m]{\frac{g(T-k)\alpha(T-k)E(T-k)/\delta}{\lambda}} \\ &\quad + \sum q_i \sqrt[m]{\frac{(1-\alpha(T-k))E(T-k)/\delta}{\lambda}} \\ &\quad \times\left(g_i^{\frac{1}{m-1}}+Q(T-k+1)^{\frac{m}{m-1}}\right)^{\frac{m-1}{m}} \end{aligned} \quad (49)$$

By differentiating with respect to $\alpha(T-k)$ and equating to zero, we obtain:

$$\alpha^*(T-k) = \frac{g(T-k)^{\frac{1}{m-1}}}{g(T-k)^{\frac{1}{m-1}}+Q(T-k)^{\frac{m}{m-1}}}, \quad (50)$$

where

$$Q(T-k) = \sum q_i\left(g_i^{\frac{1}{m-1}}+Q(T-k+1)^{\frac{m}{m-1}}\right)^{\frac{m-1}{m}} \quad (51)$$

Hence, (27) and (28) hold by induction. For the last part, let us calculate $V(E(T-k),g(T-k))$

$$\begin{aligned} &V(E(T-k),g(T-k)) \\ &= \sqrt[m]{\frac{g(T-k)g(T-k)^{\frac{1}{m-1}}E(T-k)/\delta}{\lambda(g(T-k)^{\frac{1}{m-1}}+Q(T-k)^{\frac{m}{m-1}})}} \\ &\quad + \sum q_i \sqrt[m]{\frac{Q(T-k)^{\frac{m}{m-1}}E(T-k)}{\lambda(g(T-k)^{\frac{1}{m-1}}+Q(T-k)^{\frac{m}{m-1}})}} \\ &\quad \times\left(g_i^{\frac{1}{m-1}}+Q(T-k+1)^{\frac{m}{m-1}}\right)^{\frac{m-1}{m}} \\ &= \sqrt[m]{\frac{E(T-k)/\delta}{\lambda(g(T-k)^{\frac{1}{m-1}}+Q(T-k)^{\frac{m}{m-1}})}} \\ &\quad \times(g(T-k)^{\frac{1}{m-1}}+Q(T-k)^{\frac{m}{m-1}}) \\ &= \sqrt[m]{\frac{E(T-k)/\delta}{\lambda}}\left(g(T-k)^{\frac{1}{m-1}}+Q(T-k)^{\frac{m}{m-1}}\right)^{\frac{m-1}{m}}. \end{aligned} \quad (52)$$

Thus, (29) also holds by induction.

APPENDIX C
PROOF OF THEOREM 3

The proof is by induction. We will show that the result of the theorem is true for $J_t(E(t))$ for all $t=1,\ldots,T-1$. The result of the theorem is verified for $t=T-1$ in (37). Let us assume that the theorem holds for $t+1$, i.e., if $E(t+1)\geq\gamma(t+1)$, it is optimal to stop the EH process, where $\gamma(t+1)$ is the solution to the following equation,

$$\sum q_i \sqrt[m]{1+\frac{e_i}{\gamma(t+1)}} = \frac{Q(t)}{Q(t+1)} \quad (53)$$

At time slot t we have:

$$J_t(E(t)) = \max\left(\sqrt[m]{\frac{E(t)}{\delta\lambda}}Q(t-1), \mathbb{E}(J_{t+1}(E(t+1))|E(t)\right) \quad (54)$$

First, let us assume that $E(t)\geq\gamma(t+1)$. Since $E(t+1)\geq E(t)$, it readily follows that $E(t+1)\geq\gamma(t+1)$. Thus, we have

$$\mathbb{E}(J_{t+1}(E(t+1))|E(t)) = \sum q_i \sqrt[m]{\frac{E(t)+e_i}{\delta\lambda}}Q(t)) \quad (55)$$

Hence,

$$J_t(E(t)) = \max\left(\sqrt[m]{\frac{E(t)}{\delta\lambda}}Q(t-1), \sum q_i \sqrt[m]{\frac{E(t)+e_i}{\delta\lambda}}Q(t)\right) \quad (56)$$

Since, $Q(t-1) > Q(t)$, if $E(t) \geq \gamma(t)$, then it is optimal to stop the EH process, and $\gamma(t)$ is the solution of,

$$\sum q_i \sqrt[m]{1+\frac{e_i}{\gamma(t)}} = \frac{Q(t-1)}{Q(t)}. \quad (57)$$

Note that the left hand side of (57) is strictly decreasing with respect to $\gamma(t)$ and its range is $(1\ \infty)$. Since $\frac{Q(t-1)}{Q(t)} > 1$ is proved in Lemma 1, there is a unique solution for $\gamma(t)$ satisfying (57). Thus, if $E(t) \geq \gamma(t+1)$, then the theorem is also true for case k. In the following, we will generalize the proof for any value of $E(t)$. Note that if $\gamma(t) > \gamma(t+1)$, then the proof will include any $E(t)$. Because, if $E(t) \geq \gamma(t)$, then,

$$E(t+1) \geq E(t) \geq \gamma(t) > \gamma(t+1), \quad (58)$$

and (55) will hold. Using the results of Lemma 2 we have

$$\sum q_i \sqrt[m]{1+\frac{e_i}{\gamma(t)}} < \sum q_i \sqrt[m]{1+\frac{e_i}{\gamma(t+1)}} \quad (59)$$

Hence, $\gamma(t) > \gamma(t+1)$, and the theorem holds.

REFERENCES

[1] X. Lu, P. Wang, D. Niyato, D. I. Kim, and Z. Han, "Wireless networks with RF energy harvesting: A contemporary survey," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 757–789, 2nd Quart., 2015.
[2] M. S. I. Khan, J. Mišić, and V. B. Misic, "A polling MAC with reliable RF recharging of sensor nodes," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2015, pp. 831–836.
[3] Y.-P. Hong, T.-C. Hsu, and P. Chennakesavula, "Wireless power transfer for distributed estimation in wireless passive sensor networks," *IEEE Trans. Signal Process.*, vol. 64, no. 20, pp. 5382–5395, Oct. 2016.
[4] G. Pan, H. Lei, Y. Yuan, and Z. Ding, "Performance analysis and optimization for SWIPT wireless sensor networks," *IEEE Trans. Commun.*, vol. 65, no. 5, pp. 2291–2302, May 2017.
[5] Y. Zhang, H. Pflug, H. J. Visser, and G. Dolmans, "Wirelessly powered energy autonomous sensor networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2014, pp. 2444–2449.
[6] A. Mayberry, P. Hu, B. Marlin, C. Salthouse, and D. Ganesan, "iShadow: Design of a wearable, real-time mobile gaze tracker," in *Proc. 12th Annu. Int. Conf. Mobile Syst., Appl., Services (MobiSys)*, 2014, pp. 82–94.
[7] M. Rostami, J. Gummeson, A. Kiaghadi, and D. Ganesan, "Polymorphic radios: A new design paradigm for ultra-low power communication," in *Proc. Conf. ACM Special Interest Group Data Commun. (SIGCOMM)*, 2018, pp. 446–460.
[8] 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.
[9] X. Di, K. Xiong, P. Fan, H.-C. Yang, and K. B. Letaief, "Optimal resource allocation in wireless powered communication networks with user cooperation," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 7936–7949, Dec. 2017.
[10] D. Xu and Q. Li, "Joint power control and time allocation for wireless powered underlay cognitive radio networks," *IEEE Wireless Commun. Lett.*, vol. 6, no. 3, pp. 294–297, Jun. 2017.
[11] A. Biason and M. Zorzi, "Battery-powered devices in WPCNs," *IEEE Trans. Commun.*, vol. 65, no. 1, pp. 216–229, Jan. 2017.
[12] 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. Int. Symp. Modeling Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)*, May 2016, pp. 1–7.
[13] M. A. Abd-Elmagid, A. Biason, T. ElBatt, K. G. Seddik, and M. Zorzi, "Non-orthogonal multiple access schemes in wireless powered communication networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.
[14] M. A. Abd-Elmagid, T. ElBatt, and K. G. Seddik, "Optimization of wireless powered communication networks with heterogeneous nodes," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2015, pp. 1–7.
[15] C. Zhong, G. Zheng, Z. Zhang, and G. K. Karagiannidis, "Optimum wirelessly powered relaying," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1728–1732, Oct. 2015.
[16] H. Chen, Y. Li, J. L. Rebelatto, B. F. Uchôa-Filho, and B. Vucetic, "Harvest-then-cooperate: Wireless-powered cooperative communications," *IEEE Trans. Signal Process.*, vol. 63, no. 7, pp. 1700–1711, Apr. 2015.
[17] Y. Gu, H. Chen, Y. Li, and B. Vucetic, "An adaptive transmission protocol for wireless-powered cooperative communications," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2015, pp. 4223–4228.
[18] D. Xu and H. Zhu, "Outage minimized resource allocation for multiuser OFDM systems with SWIPT," *IEEE Access*, vol. 7, pp. 79714–79725, 2019.
[19] H. Ju and R. Zhang, "User cooperation in wireless powered communication networks," in *Proc. IEEE Global Commun. Conf.*, Dec. 2014, pp. 1430–1435.
[20] M. Zhong, S. Bi, and X. Lin, "User cooperation for enhanced throughput fairness in wireless powered communication networks," in *Proc. 23rd Int. Conf. Telecommun. (ICT)*, May 2016, pp. 1–6.
[21] S. Bi and Y. J. Zhang, "Computation rate maximization for wireless powered mobile-edge computing with binary computation offloading," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 4177–4190, Jun. 2018.
[22] C. You and K. Huang, "Wirelessly powered mobile computation offloading: Energy savings maximization," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2015, pp. 1–6.
[23] F. Wang, J. Xu, X. Wang, and S. Cui, "Joint offloading and computing optimization in wireless powered mobile-edge computing systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1784–1797, Dec. 2017.
[24] F. Wang, J. Xu, X. Wang, and S. Cui, "Joint offloading and computing optimization in wireless powered mobile-edge computing systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–6.
[25] O. Ozel, K. Tutuncuoglu, J. Yang, S. Ulukus, and A. Yener, "Transmission with energy harvesting nodes in fading wireless channels: Optimal policies," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 8, pp. 1732–1743, Sep. 2011.
[26] J. Yang and S. Ulukus, "Optimal packet scheduling in an energy harvesting communication system," *IEEE Trans. Commun.*, vol. 60, no. 1, pp. 220–230, Jan. 2012.
[27] Z. Wang, V. Aggarwal, and X. Wang, "Power allocation for energy harvesting transmitter with causal information," *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 4080–4093, Nov. 2014.
[28] M. L. Ku, Y. Chen, and K. J. R. Liu, "Data-driven stochastic models and policies for energy harvesting sensor communications," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 8, pp. 1505–1520, Aug. 2015.
[29] R. Ma and W. Zhang, "Optimal power allocation for energy harvesting communications with limited channel feedback," in *Proc. IEEE Global Conf. Signal Inf. Process. (GlobalSIP)*, Atlanta, GA, USA, Dec. 2014, pp. 193–197.
[30] M. R. Zenaidi, Z. Rezki, and M. S. Alouini, "Performance limits of online energy harvesting communications with noisy channel state information at the transmitter," *IEEE Access*, vol. 5, pp. 1239–1249, 2017.
[31] W. Du, J. C. Liando, H. Zhang, and M. Li, "Pando: Fountain-enabled fast data dissemination with constructive interference," *IEEE/ACM Trans. Netw.*, vol. 25, no. 2, pp. 820–833, Apr. 2017.
[32] M. S. H. Abad, O. Ercetin, and D. Gündüz, "Channel sensing and communication over a time-correlated channel with an energy harvesting transmitter," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 1, pp. 114–126, Oct. 2017.
[33] M. S. H. Abad, O. Ercetin, T. E. Batt, and M. Nafie, "SWIPT using hybrid ARQ over time varying channels," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 4, pp. 1087–1100, Aug. 2018.
[34] B. T. Bacinoglu, E. Uysal-Biyikoglu, and C. E. Koksal, "Finite-horizon energy-efficient scheduling with energy harvesting transmitters over fading channels," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 6105–6118, Sep. 2017.

[35] M. S. H. Abad and O. Ercetin, "Finite horizon throughput maximization for a wirelessly powered device over a time varying channel," in *Proc. IEEE Globecom Workshops*, Dec. 2018, pp. 1–6.
[36] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998.
[37] W. Zhang, Y. Wen, K. Guan, D. Kilper, H. Luo, and D. O. Wu, "Energy-optimal mobile cloud computing under stochastic wireless channel," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4569–4581, Sep. 2013.
[38] S.-W. Ko, K. Huang, S.-L. Kim, and H. Chae, "Live prefetching for mobile computation offloading," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 3057–3071, May 2017.
[39] J. Lee and N. Jindal, "Energy-efficient scheduling of delay constrained traffic over fading channels," *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 1866–1875, Apr. 2009.
[40] M. Zafer and E. Modiano, "Delay-constrained energy efficient data transmission over a wireless fading channel," in *Proc. Inf. Theory Appl. Workshop*, Jan. 2007, pp. 289–298.
[41] C. You, Y. Zeng, R. Zhang, and K. Huang, "Resource management for asynchronous mobile-edge computation offloading," in *Proc. IEEE Int. Conf. Commun. (ICC) Workshops*, May 2018, pp. 1–6.
[42] C. You, Y. Zeng, R. Zhang, and K. Huang, "Asynchronous mobile-edge computation offloading: Energy-efficient resource management," *IEEE Trans. Wireless Commun.*, vol. 17, no. 11, pp. 7590–7605, Sep. 2018.
[43] D. J. Russo, B. V. Roy, A. Kazerouni, I. Osband, and Z. Wen, "A tutorial on thompson sampling," *Found. Trends Mach. Learn.*, vol. 11, no. 1, pp. 1–96, 2017.
[44] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, 2nd ed. Belmont, MA, USA: Athena Scientific, 2000.
[45] C. P. Robert and G. Casella, *Monte Carlo Statistical Methods* (Springer Texts in Statistics), 2nd ed. New York, NY, USA: Springer-Verlag, 2004.
[46] X. Li, C. You, S. Andreev, Y. Gong, and K. Huang, "Wirelessly powered crowd sensing: Joint power transfer, sensing, compression, and transmission," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 2, pp. 391–406, Feb. 2019.
[47] D. Xu and Q. Li, "Cooperative resource allocation in cognitive wireless powered communication networks with energy accumulation and deadline requirements," *Sci. China Inf. Sci.*, vol. 62, no. 8, p. 82302, Jul. 2019.
[48] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in *Proc. 25th Annu. Conf. Learn. Theory*, Jun. 2012, pp. 39.1–39.26.
[49] E. Kaufmann, N. Korda, and R. Munos, "Thompson sampling: An asymptotically optimal finite-time analysis," in *Algorithmic Learning Theory*. Berlin, Germany: Springer, 2012, pp. 199–213.
[50] S. Agrawal and N. Goyal, "Further optimal regret bounds for thompson sampling," in *Proc. 16th Int. Conf. Artif. Intell. Statist. (AISTATS)*, Apr. 2013, pp. 99–107.

MEHDI SALEHI HEYDAR ABAD received the B.S. degree in electrical engineering from IUST, Tehran, Iran, in 2012, and the M.S. degree in electrical and electronics engineering from Sabanci University, Istanbul, Turkey, in 2015, where he is currently pursuing the Ph.D. degree. He was a Visiting Researcher with The Ohio State University, Columbus, OH, USA. His research interests include mathematical modeling of communication systems, stochastic optimization, and green communication networks. He was a recipient of the Best Paper Award at the 2016 IEEE Wireless Communications and Networking Conference (WCNC).

OZGUR ERCETIN received the B.S. degree in electrical and electronics engineering from Middle East Technical University, Ankara, Turkey, in 1995, and the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, MD, USA, in 1998 and 2002, respectively. Since 2002, he has been with the Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul, Turkey. He was a Visiting Researcher with HRL Laboratories, Malibu, CA, USA, Docomo USA Labs, Palo Alto, CA, USA, The Ohio State University, Columbus, OH, USA, Carleton University, Ottawa, ON, Canada, and the Université du Québec à Montréal, Montreal, QC, Canada. His research interests include computer and communication networks with an emphasis on fundamental mathematical models, architectures and protocols of wireless systems, and stochastic optimization.