Communication over a Gilbert-Elliot Channel with an Energy Harvesting Transmitter

Abstract

Communication over a wireless point-to-point link, exhibiting time correlated behavior, is studied for an energy harvesting (EH) transmitter. At the beginning of the time slot, the EH transmitter, without knowing the realization of the current channel state, has three possible actions: *i*) deferring the transmission to save its energy for future use, *ii*) transmitting at a low rate of R_1 bits with guaranteed successful delivery of the message *iii*) transmitting at a high rate rate of R_2 bits, and *iv*) sensing the channel to reveal the channel state by consuming a portion of its energy and transmission time, followed by transmission at a reduced rate consuming the remainder of the energy unit, if the channel is in the good state. We aim to maximize the total expected discounted number of bits transmitted over an infinite time horizon. This problem can be formulated as a partially observable Markov decision process (MDP) which is then converted to an ordinary MDP by introducing a belief on the channel state, and the optimal policy is shown to exhibit a threshold behavior on the belief state, with battery-dependent threshold values. Optimal threshold values and corresponding optimal performance are characterized through numerical simulations and it is shown that having the sensing action and intelligently using it to track the channel state improves the achievable long-term throughput significantly.

I. Introduction

Due to the tremendous increase in the number of battery-powered wireless communication devices over the past decade, replenishing the batteries of these devices by harvesting energy from natural resources has become an important research area [1]. Transmitters may harvest energy via wind turbines, photovoltaic cells, thermoelectric generators, or from mechanical vibrations through piezoelectric or electromagnetic technology [2]. Regardless of which type of EH device and natural energy source is employed, a main concern is the stochastic nature of the EH process driving the wireless communications. The associated battery recharging process can be modeled either as a continuous, or a discrete, [3], [4] stochastic process.

We consider a wireless point-to-point link with a transmitter equipped with a finite-capacity battery fed by an EH device. At each time slot, a unit of energy is harvested by the transmitter according to a binary random process independent over time[1]. We assume that the transmitter can accurately observe the current energy level of the battery, and it has the knowledge of the statistics of the EH process. The wireless channel is time-varying and has memory across time. The channel memory is modeled with a finite state Markov chain [5], where the next channel state depends only on the current state. A convenient and often-employed simplification of the Markov model is a two state Markov chain, known as the Gilbert-Elliot channel [6]. This model assumes that the channel can be either in a *good* or a *bad* state. We assume that, by spending exactly one unit of energy from its battery, in the good state, transmitter may transmit R_2 bits of information reliably, while in the bad state it may transmit at a lower rate of R_1 bits.

In this work, differently from most of the literature on EH systems, we take into account the energy cost of acquiring channel state information (CSI). At the beginning of each time slot, without knowing the current CSI, EH transmitter has three possible actions: *i*) deferring the transmission to save its energy for future use, *ii*) transmitting at a low rate of R_1 bits with guaranteed successful delivery of the message *iii*) transmitting at a high rate rate of R_2 bits, and *iv*) sensing the channel to reveal the channel state by consuming a portion of its energy and transmission time, followed by transmission at a reduced rate consuming the remainder of the energy unit, if the channel is in the good state. If the channel is in a bad state, the transmitter either remains silent in the rest of the time slot, saving its energy for future or transmits at a lower rate of R_1 bits by utilizing the rest of the energy unit. If the level of the battery is less than a unit of energy at the beginning of a time slot, no transmission is possible. Our objective is to maximize the total expected discounted number of bits transmitted over an infinite time horizon.

A. Related Work

Markov decision process (MDP) tools have been extensively utilized in the recent literature in solving communication problems involving EH devices. In [7] authors propose a simple single-threshold policy for a solar-powered sensor operating over a fading wireless channel. Optimality of a single-threshold policy is proven [8] when transmitting packets with importance values on EH transmitter. Problem of energy allocation for gathering and transmitting of data in an

[1] Typically, the EH process is neither memoryless nor discrete, and the energy is accumulated continuously over time. However, in order to develop the analytical model underlying this paper, we follow the common assumption in the literature [3], and assume that the continuous energy arrival is accumulated in an intermediate energy storage device to form quantas.

EH communication system is studied in [9] and [10]. The scheduling of EH transmitters with time correlated energy arrivals to optimize the long term sum throughput is investigated in [11]. The allocation of energy over a finite horizon to optimize the throughput is considered in [12], where it is assumed that either the current or the future energy and channel states are provided to the transmitter. In [13], for a Markov EH process, and static channel process a discrete power allocation problem is studied to maximize the throughput. In [14] throughput is optimized over a multiple access channel with collusions, considering spatially correlated energy arrivals at the transmitters.

In a closely related work [15], scheduling of an EH transmitter over a Gilbert-Elliot channel is considered. However, unlike our work, the transmitter [15] always has perfect CSI, obtained by sensing at every time slot, and makes a decision to defer or to transmit, based on the current CSI and battery state. Similarly, without considering the channel sensing capability, [16] addresses the problem of optimal power management for an EH sensor over a multi-state wireless channel with memory. In our work, instead, we take into account the energy cost of channel sensing which can be significant for EH transmitters. Therefore, the EH transmitter does not necessarily have perfect CSI, but keeps an updated belief of the channel state according to its past observations. Hence, the transmitter may occasionally take a third decision (in addition to defer and to transmit) of sensing the current channel state to improve its belief.

Channel sensing is an essential part of opportunistic and cognitive spectrum access. In [17], the authors investigate the problem of optimal access to a Gilbert-Elliot channel, wherein an energy-unlimited transmitter senses the channel at every time slot. In [18] channel sensing is done only occasionally. The transmitter can decide to transmit at a high or a low rate without sensing the channel; or can first sense the channel and transmit at a reduced rate due to the time spent for sensing. The energy cost of sensing is ignored in [18].

B. Organization of the paper

In Section II we explain the channel and EH process models under consideration, and elaborate on the transmission protocol. In Section III, we formulate the problem as a two state partially observable MDP (POMDP) which is then converted to a continuous-state MDP by introducing a belief state. In Section IV we show that the optimal policy is of threshold type, for which the optimal threshold values depend on the state of the battery. In Section V we present simulation

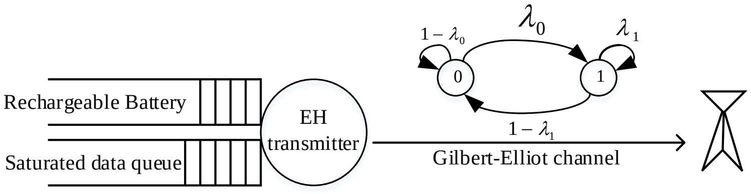


Fig. 1: System model.

results that numerically obtain the optimal threshold values and the optimal performance. In Section VI we conclude the paper and present future research directions.

II. System Model

A. Channel and energy harvesting models

Consider the communication system illustrated in Fig. 1, in which an EH transmitter communicates over a slotted Gilbert-Elliot channel. Let G_t denote the state of the channel at time slot t which is modeled as a one-dimensional Markov chain with two states: a good state denoted by 1, and a bad state denoted by 0. Channel transitions occur at the beginning of each time slot. The transition probabilities are given by $\mathbb{P}\left[G_t = 1 | G_{t-1} = 1\right] = \lambda_1$ and $\mathbb{P}\left[G_t = 1 | G_{t-1} = 0\right] = \lambda_0$. We assume that the states are positively correlated, that is, $\lambda_0 \leq \lambda_1$. The transmitter is able to transmit R_2 bits per time slot if $G_t = 1$, and $R_1 < R_2$ bits if $G_t = 0$.

A unit of energy arrives at the end of time slot t according to an independent and identically distributed (i.i.d.) Bernoulli process, denoted by E_t, with probability q, i.e., $\mathbb{P}\left[E_t = 1\right] = q$ for all t. The transmitter stores the energy packets in a battery with a storage capacity of B_{max} units of energy. We denote the state of the battery, i.e., the energy available in the battery at the beginning of time slot t, by B_t. An energy unit is consumed at each slot if the transmitter decides to transmit in that slot. A unit of energy consumed per slot includes the energy cost of sensing (if the transmitter decides to sense the channel), transmission of the message, and the reception of ACK or NAK from the receiver. We assume that the transmitter has an infinitely backlogged data queue, and thus, it always has a packet to transmit.

B. Transmission protocol

Once a transmission occurs either in an good or bad state, the receiver replies with an acknowledgment (ACK) if the transmission is successful, or with a negative acknowledgment

(NAK) if the transmission fails. Note that an ACK message informs the transmitter that the most recent state of the channel was good, whereas a NAK message informs the transmitter that the most recent state of the channel was bad.

At the beginning of each time slot, the transmitter may choose among three possible actions: *i*) defer the transmission, *ii*) transmitting at low rate, *iii*) channel sensing and *iv*) transmitting without sensing.

Defer the transmission: This action (denoted by D) corresponds to the case in which the transmitter either believes that the channel is in a bad state, or observes that its battery has low energy. This action helps the transmitter to prevent energy outage in its battery which disables the transmitter to send any data, even if it believes that the channel is in good state. If this action is chosen, there is no message exchange between the transmitter and the receiver. Hence, the receiver does not send any feedback, and therefore the transmitter cannot obtain any knowledge about the current channel state. The scenario in which the transmitter is informed about the current channel state even when it does not transmit any data packet is equivalent to the system model investigated in [15].

Transmitting at low rate: If this action is chosen (denoted by L), the transmitter behaves conservatively and transmits lower number of data bits using high redundancy coding schemes to guarantee successful delivery of its message to the destination at the cost of low data rate. According to the adopted channel model in this paper, it is possible to deliver R_1 bits, at any time slot, regardless of the channel state realization. Since the transmitter is assured about the successful delivery of its message, at the end of the transmission, the receiver does not send any feedback, and therefore the transmitter cannot obtain any knowledge about the current channel state.

Channel sensing: This action corresponds to the case in which the transmitter decides to sense the channel at the beginning of the time slot. We assume that sensing consumes a fraction $0 < \tau < 1$ of an energy unit. Sensing is carried out by the transmitter first sending a control/probing packet, to which, the receiver responds with a packet indicating the channel state. We assume that the time it takes to sense the channel is τ seconds and the transmitter consumes on average the same power as data transmission over the sensing period. Therefore, we equivalently assume for simplicity that $\tau = 1/k$ for some $k \in \mathbb{Z}^+$. In the remaining $1-\tau$ seconds, the transmitter may choose to transmit data at the same rate it would without channel sensing, which means that by the end of the time slot it transmits $(1-\tau)R$ bits.

If the channel is revealed to be in the bad state, the transmitter has to option; either it defers its transmission (denoted by *OD*) and saves the rest of the energy unit (i.e., $1-\tau$) or utilizes the rest of the energy unit to transmit at a lower rate, R_1 (denoted by *OT*). Note that thanks to the channel sensing capability, in the case of a bad state, the transmitter wastes only τ portion of a unit energy packet, and either saves the remaining energy by deferring its transmission or utilizes it by sending lower number of data bits to the destination. As we will show later in this paper, is an important advantage in EH networks with scarce energy sources.

Transmitting without sensing: This action (denoted by T) corresponds to the case when transmitter attempts to transmit R_2 bits in the current slot without sensing the channel. If the channel is in a good state, the transmission is successful and the receiver sends an ACK. Otherwise, the transmission fails, and the receiver sends a NAK. Note that at the end of the slot the transmitter has the perfect knowledge of the current channel state.

III. Partially Observable Markov Decision Process (POMDP) formulation

At the beginning of each time slot, the transmitter chooses among the four possible actions based on the state of its battery, and its belief about the channel state to maximize a long-term discounted reward to be defined shortly. Although the transmitter is perfectly aware of its battery state, it cannot directly observe the current channel state. Hence, the problem in hand becomes a partially observable Markov decision process (POMDP).

Let the state of the system at time t be denoted by $S_t = (B_t, X_t)$. We define the *belief* of the transmitter at time slot t, denoted by X_t, as the conditional probability that the channel is in the good state at the beginning of the current slot, i.e., $X_t = \mathbb{P}\left[G_t = 1 | \mathscr{H}_t\right]$, given the history $\mathscr{H}_t$, where $\mathscr{H}_t$ represents all the past actions and observations of the transmitter up to slot t. The transmitter's belief constitutes a sufficient statistic to characterize its optimal actions [19]. Note that with this definition of the state, the POMDP problem is converted into a MDP with an uncountable state space $[0, \tau, 2\tau, \ldots, B_{max}] \times [0,\ 1]^2$.

A transmission policy π describes a set of rules that dictates which action to take depending on the history. Let $V^{\pi}(b, p)$ be the expected infinite-horizon discounted reward with initial state $S_0 = (b,\ \mathbb{P}\left[G_0 = 1 | \mathscr{H}_0\right] = p)$ under policy π with discount factor $\beta \in [0,\ 1)$. The use of the expected discounted reward allows us to obtain a tractable solution, and one can gain insights

[2] Note that since sensing without transmission is possible, i.e., consuming only τ fraction of the energy unit, the battery can take fraction of units as states.

into the optimal policy for the average reward when β is close to 1. It is also discussed in [5] that β can be interpreted as the probability that a particular user is allowed to use the channel, or as the probability of the transmitter to remain active at each time slot as in [20]. For an initial belief p the expected discounted reward has the following expression

$$V^{\pi}(b,\ p) = \mathbb{E}\left[\sum_{t=0}^{\infty} \beta^t R(S_t, A_t)|S_0 = (b,\ p)\right], \tag{1}$$

where t is the time index, $A_t \in \{D, O, T\}$ is the action chosen at time t, and $R(S_t, A_t)$ is the expected reward acquired when action A_t is taken at state S_t. The expectation in (1) is over state sequence distribution induced by the given transmission policy π. The expected reward when action A_t is chosen at state S_t is given as follows:

$$R(S_t, A_t) = \begin{cases} X_t R_2 & \text{if } A_t = T \text{ and } B_t \geq 1, \\ R_1 & \text{if } A_t = L \text{ and } B_t \geq 1, \\ (1-\tau)X_t R_2 & \text{if } A_t = OD \text{ and } B_t \geq 1, \\ (1-\tau)[(1-X_t)R_1 + X_t R_2] & \text{if } A_t = OT \text{ and } B_t \geq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Since at least one energy unit is required for transmission, if the battery state is less than one unit, the reward becomes zero. Hence, in explaining the expected reward function in (13), we consider actions when the battery state is greater than or equal to one. If the action of transmitting without sensing is chosen, R_2 bits are transmitted successfully if the channel is in a good state, and 0 bits if the channel is in a bad state. Since the belief, X_t, represents the probability of the channel being in a good state, the expected reward is given by $X_t R_2$. Since it is guaranteed that transmitting at low rate is always successful, the expected reward for this action is R_1. If the action of channel sensing is chosen, τ fraction of energy unit is spent sensing the channel with the remaining energy either being used for transmission or it is stored in the battery. If the channel is sensed to be in a good state, $(1-\tau)R_2$ bits are transmitted successfully. If the channel is sensed to be in a bad state, the transmitter either remains silent by saving the rest of the energy unit and receives no rewards or utilizes the rest of the energy unit and receives $(1-\tau)R_1$ in the rest of the time slot. Thus the expected reward by choosing action OD is $(1-\tau)X_t R_2$ and the expected reward by choosing OT is $(1-\tau)[(1-X_t)R_1 + X_t R_2]$. Finally, if the action of deferring the transmission is taken the transmitter neither senses the channel nor transmits, so the reward is zero.

Define the value function $V(b,\ p)$ as

$$V(b,\ p) = \max_{\pi} V^{\pi}(b,\ p)$$
$$\text{for all } b \in [0, \tau, 2\tau, \ldots, B_{max}] \text{ and } p \in [0,\ 1]. \tag{3}$$

It is well known that the optimal value of the infinite-horizon expected reward can be achieved by a stationary policy, i.e., there exists a stationary policy π^* such that $V(b,\ p) = V^{\pi^*}(b,\ p)$ [21]. The value function $V(b,\ p)$ satisfies the Bellman equation

$$V(b,\ p) = \max_{A \in \{D,O,T\}} \{V_A(b,\ p)\}, \tag{4}$$

where $V_A(b,\ p)$ is the action-value function, defined as the expected infinite-horizon discounted reward acquired by taking action A when the state is $(b,\ p)$, and is given by

$$\begin{aligned} V_A(b,\ p) =& R((b,\ p),A) \\ &+ \beta \mathbb{E}_{(\acute{b},\ \acute{p})} \left[V(\acute{b},\ \acute{p}) | S_0 = (b,\ p), A_0 = A \right], \end{aligned} \tag{5}$$

where $(\acute{b},\ \acute{p})$ denotes the next state when action A is chosen at state $S_0 = (b,\ p)$. The expectation in (5) is over the distribution of possible next states. In the following, we define and explain the value function $V_A(b,\ p)$, and how the system state evolves for each action.

Defer the transmission: If this action is taken, since there is no transmission, there is no ACK or NAK from the receiver, and thus, the transmitter does not learn the state of the channel. Therefore the next belief is obtained as the probability of finding the channel in a good state given the current belief state. If the transmitter had a belief $X_t = p$ at time slot t, after taking action D, its belief at the beginning of the next slot is updated as

$$J(p) = \lambda_0(1-p) + \lambda_1 p. \tag{6}$$

In every time slot, a unit of energy is harvested with probability q. Thus, after taking action D, the value function evolves as follows:

$$V_D(b,\ p) = \beta \left[qV\left(\min\{b+1, B_{max}\},\ J(p)\right) + (1-q)V\left(b,\ J(p)\right)\right]. \tag{7}$$

Note that the term $\min\{b+1, B_{max}\}$ is used to ensure that the battery state does not exceed the battery capacity, B_{max}.

Transmitting at low rate: Since transmission costs one unit of energy, this action can be taken if $b \geq 1$[3]. If this action is taken , the transmission will be successful independent of the channel state. Hence, there is no ACK or NAK from the receiver and as a result the transmitter does not learn the state of the channel. Similar to action D, the next belief of the channel state will be updated as $J(p) = \lambda_0(1-p) + \lambda_1 p$ and the value function corresponding to this action will be updated as follow:

$$V_L(b,\ p) = R_1 + \beta \left[qV\left(\min\{b, B_{max}\},\ J(p)\right) + (1-q)V\left(b-1,\ J(p)\right)\right]. \tag{8}$$

Channel sensing: For this action, depending on the battery state, two scenarios are possible. If $b \geq 1$ and EH decides to sense the channel, then it consumes τ fraction of energy to first sense the channel and obtain the current channel state. Based on the outcome of the channel sensing, if the channel is found to be in a good state, $(1-\tau)$ units of energy is used to transmit $(1-\tau)R_2$ bits. Also, the belief state is updated as λ_1 for the next time slot.

On the other hand, if the outcome of the channel sensing reveals the channel to be in a bad state, then the transmitter either defers its transmission (OD), and saves $(1-\tau)$ units of energy for possible future transmissions or utilizes $(1-\tau)$ units of energy and transmits at the rate of R_1 (OT). Also, the channel belief is updated as λ_0 for the next time slot. Based on the aforementioned discussion, for $b \geq 1$ the evolution of the value function can be written as:

$$\begin{aligned} V_{OD}(b,\ p) =& p\left[(1-\tau)R_2 + \beta\left(qV(b,\ \lambda_1) + (1-q)V(b-1,\ \lambda_1)\right)\right] \\ &+ (1-p)\beta\left[qV(\min\{b-\tau+1, B_{max}\},\ \lambda_0) + (1-q)V(b-\tau,\ \lambda_0)\right]. \end{aligned} \tag{9}$$

$$\begin{aligned} V_{OT}(b,\ p) =& p\left[(1-\tau)R_2 + \beta\left(qV(b,\ \lambda_1) + (1-q)V(b-1,\ \lambda_1)\right)\right] \\ &+ (1-p)\left[(1-\tau)R_1\beta\left(V(b,\ \lambda_0) + (1-q)V(b-1,\ \lambda_0)\right)\right] \end{aligned} \tag{10}$$

If $\tau \leq b < 1$, then transmission is not possible since transmission requires at least one unit of energy. However, it is still possible to sense the channel, since it only requires τ fraction of

[3] Note that we are aware that in the generic MDP formulation, in every state, we should have the same set of actions. We can re-define the reward function by assigning $-\infty$ reward for those actions that are not possible to be taken in specific states to account for this issue. For the ease of comprehension, we chose to present the formulation in this manner.

energy. This may happen when EH node believes that learning the channel state will help its decision in the future. Thus for $\tau \leq b < 1$, the value function evolves as:

$$\begin{aligned} V_{OD}(b,\ p) = V_{OT}(b,\ p) =&\beta \left[\, qpV(b-\tau+1,\ \lambda_1) \right. \\ &+ q(1-p)V(b-\tau+1,\ \lambda_0) + (1-q)pV(b-\tau,\ \lambda_1) \\ &\left. + (1-q)(1-p)V(b-\tau,\ \lambda_0) \,\right] \end{aligned} \tag{11}$$

Note that in (11), only sensing part of the actions *OD* and *OT* is feasible since it is obviously impossible to transmit any data when $\tau \leq b < 1$.

Transmitting without sensing: This action can only be chosen if the battery state is greater than or equal to one, i.e., $b \geq 1$. Under this action, the transmitter transmits regardless of the actual state of the channel, costing one unit of energy. If the channel is in the good state, R bits are successfully delivered to the receiver, and the receiver sends back an ACK. Otherwise, the channel is in the bad state, so the transmission fails, and the receiver sends back a NAK. Meanwhile, the channel is in a good state with probability p, i.e., the current belief state, and the belief in the next time slot will be λ_1. Also the channel is in a bad state with probability $1-p$ and the belief in the next time slot will be λ_0. Hence, the value function evolves as:

$$\begin{aligned} V_T(b,\ p) =&p\left[R_2 + \beta\left(qV(b,\ \lambda_1) + (1-q)V(b-1,\ \lambda_1)\right)\right] \\ &+ (1-p)\beta\left[V(b,\ \lambda_0) + (1-q)V(b-1,\ \lambda_0)\right] \end{aligned} \tag{12}$$

IV. Structure of The Optimal Policy

For the given set of actions $\{D, L, OD, OT, A\}$, it is possible to prove that the optimal policy has a threshold type structure on the belief state. The belief state set, i.e., the interval [0, 1] can be portioned into subsets of the interval [0, 1], and then each subset is assigned to an action. However, since the actions D and L are not necessarily convex with respect to the belief p, mathematically speaking, the partitioning process can be done in an infinitely many way. We bypass this problem by assuming that it is not possible to transmit any bits whenever the channel is in a bad state, i.e., $R_1 = 0$ and $R_2 = R$. Note that in this case action L does not exist and channel sensing action reduces to only *OD* which for simplicity we denote by *O* afterwards.

With the aforementioned assumptions, the new expected reward function can be written as follows:

$$R(S_t, A_t) = \begin{cases} X_t R & \text{if } A_t = T \text{ and } B_t \geq 1, \\ (1-\tau) X_t R & \text{if } A_t = O \text{ and } B_t \geq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

We prove that the optimal policy has a threshold type structure on the belief state. First, we need to prove some of the properties of the value function. We begin with establishing the convexity of the optimal value function with respect to the belief state.

Lemma 1. *For any given $b \geq 0$, V(b, p) is convex in p.*

Proof. The proof is given in Appendix A. □

In the following lemma, we show that the value function is a non decreasing function of battery state, b. This lemma provides the intuition why deferring or sensing actions are advantageous in some states. The incentive of taking these actions is that the value function transitions into higher values without consuming any energy.

Lemma 2. *Given any belief p, $V(b_1, p) \geq V(b_0, p)$ when $b_1 > b_0$.*

Proof. The proof is given in Appendix B. □

The next result states that the value function is also non-decreasing with respect to the belief state, p.

Lemma 3. *For a fixed battery state b, if $p_1 > p_0$ then $V(b, p_1) \geq V(b, p_0)$.*

Proof. The proof is given in Appendix C. □

Finally, Theorem 1 below shows that the optimal solution of the problem is a threshold policy with two or three thresholds depending on the system parameters. The threshold values depend on the state of the battery.

Theorem 1. *Let $p \in [0,\ 1]$ and $b \geq 0$, there are thresholds $0 \leq \rho_1(b) \leq \rho_2(b) \leq \rho_3(b) \leq 1$, all* *functions of the battery state, b, such that for* $b \geq 1$

$$\pi^*(b,\ p) = \begin{cases} D, & \textit{if}\ \ 0 \leq p \leq \rho_1(b)\ \textit{or}\ \rho_2(b) \leq p \leq \rho_3(b) \\ O, & \textit{if}\ \ \rho_1(b) \leq p \leq \rho_2(b), \\ A, & \textit{if}\ \ \rho_3(b) \leq p \leq 1, \\ & \end{cases} \tag{14}$$

and for $\tau \leq b < 1$*,*

$$\pi^*(b,p) = \begin{cases} D, & \textit{if}\ \ 0 \leq p \leq \rho_1(b)\ \textit{or}\ \rho_2(b) \leq p \leq 1, \\ O, & \textit{if}\ \ \rho_1(b) \leq p \leq \rho_2(b). \end{cases} \tag{15}$$

Proof. For any $b \in \{0, \tau, 2\tau, \ldots, B_{max}\}$, We define the following sets:

$$\Phi_T^b \triangleq \{p \in [0,1] : V(b,p) = V_T(b,p)\}\,,\ \text{for}\ T \in \{D,O,A\}\,. \tag{16}$$

Note that, given any battery state $b \geq 0$, Φ_T^b characterizes the set of belief states for which it is optimal to choose the action T. It is easy to see that for $b = 0$, $V(b,p) = V_D(b,p)$, and hence $\Phi_D^0 = [0,\ 1]$, and $\Phi_D^0 = \Phi_D^0 = \varnothing$. First, we consider battery states $\tau \leq b < 1$. We will prove that for any $\tau \leq b < 1$, Φ_O^b is convex, which implies the structure of the optimal policy in (15). Let $p_1,\ p_2 \in \Phi_O^b$, and $a \in [0,\ 1]$. We have

$$\begin{aligned} V(b, ap_1 + (1-a)p_2) &\leq aV(b,p_1) + (1-a)V(b,p_2), && (17) \\ &= aV_O(b,p_1) + (1-a)V_O(b,p_2), && (18) \\ &= V_O(b, ap_1 + (1-a)p_2), && (19) \\ &\leq V(b, ap_1 + (1-a)p_2), && (20) \end{aligned}$$

where (17) follows from Lemma 1; (18) is due to the fact that $p_1,\ p_2 \in \Phi_O^b$; (19) follows from the linearity of V_O in p; and (20) holds due to the definition of $V(b,\ p)$. Consequently, $V(ap_1 + (1-a)p_2) = V_O(ap_1 + (1-a)p_2)$, and it follows that $ap_1 + (1-a)p_2 \in \Phi_O^b$, which, in turn, proves the convexity of Φ_O^b. Note also that $p = 0$ and $p = 1$ both belong to Φ_D^b for all $0 \leq b < 1$. Hence, for a given battery state $0 \leq b < 1$, either $\Phi_O^b = \varnothing$, or there exists $0 < \rho_1(b) \leq \rho_2(b) < 1$ such that $\Phi_O^b = [\rho_1(b), \rho_2(b)]$. Consequently, we have $\Phi_D^b = [0, \rho_1(b)) \cup (\rho_2(b), 1]$.

Next, consider $1 \leq b \leq B_{max}$. We will prove that Φ_A^b and Φ_O^b are both convex, which implies the structure of the optimal policy in (14). Let $p_1,\ p_2 \in \Phi_A^b$ and $a \in [0,\ 1]$. Similar to (17)-(20)

we can argue

$$\begin{aligned} V(b, ap_1+(1-a)p_2) &\le aV(b,p_1)+(1-a)V(b,p_2), \\ &= aV_A(b,p_1)+(1-a)V_A(b,p_2), \\ &= V_A(b, ap_1+(1-a)p_2), \\ &\le V(b, ap_1+(1-a)p_2). \end{aligned} \tag{21}$$

Consequently, $V(ap_1+(1-a)p_2) = V_A(ap_1+(1-a)p_2)$; and hence $ap_1+(1-a)p_2 \in \Phi_A^b$, which proves the convexity of Φ_A^b. Since it is always optimal to act aggressively if the channel is in the good state (see [15]) $1 \in \Phi_A^b$, and since the convex subsets of the real line are intervals, there exists $\rho_3(b) \in (0,1]$ such that $\Phi_A^b = [\rho_3(b), 1]$. Using the same technique we can prove that Φ_O^b is convex, and hence, there exists $0 < \rho_1(b) \le \rho_2(b) < 1$ such that $\Phi_O^b = [\rho_1(b), \rho_2(b)]$. As a result the remaining segments belong to action D, and we have $\Phi_D = [0, \rho_1(b)) \cup (\rho_2(b), \rho_3(b))$. □

Theorem 1 proves that at any battery state $b \ge 1$, at most three threshold values are sufficient to characterize the optimal policy; whereas two thresholds suffice for $0 \le b < 1$. However the optimal policy can even be simpler for some battery states and some instances of the problem as it is possible to have $\rho_2(b) = \rho_3(b)$, or even $\rho_1(b) = \rho_2(b) = \rho_3(b)$. Due to the non-convexity of action D, the structure of the optimal policy may result in four different regions even though there are only three possible actions. This may be counter intuitive for the reader since deferring the transmission should not be advantageous when the belief is relatively high. Nevertheless, in Section V, we demonstrate that in some cases it is indeed optimal to have a three threshold policy.

V. Numerical Results

In this section we use numerical techniques to characterize the optimal policy, and evaluate its performance. We utilize the value iteration algorithm to calculate the optimal value function. We numerically identify the thresholds for the optimal policy for different scenarios. We also evaluate the performance of the optimal policy, and compare it with some alternative policies in terms of throughput.

A. Optimal policy evaluation

In the following, we assume that $B_{max} = 5$, $\tau = 0.2$, $\beta = 0.98$, $\lambda_1 = 0.9$, $\lambda_0 = 0.6$, $R = 3$ and $q = 0.1$. The optimal policy is evaluated using the value iteration algorithm. In Fig. 2 each state $(b,\ p)$ is illustrated with a different color corresponding to the optimal policy at that state. In the figure, the areas highlighted with blue color correspond to those states at which deferring the transmission is optimal, green areas correspond to the states at which transmitting opportunistically is optimal, and finally yellow areas correspond to the states for which transmitting without sensing is optimal. As seen in Fig. 2 any of the three policies (one, two, or three threshold policies) may be optimal depending on the level of the battery state. For example, when the battery state is $b = 2$, one-threshold policy is optimal. The transmitter defers transmission up to a belief of state of $p = 0.8$ and starts transmitting aggressively beyond this value. For no value of the belief state it opts for sensing the channel. On the other hand, when the battery state is 3.8, two-threshold policy is optimal, and when the battery state is 2.8, three-threshold policy is optimal. Considering the low probability of energy arrivals $(q = 0.1)$ and the relative high cost of sensing $(\tau = 0.2)$, it is interesting to notice that the transmitter senses the channel even when its battery state is below the transmission threshold, i.e., $b < 1$.

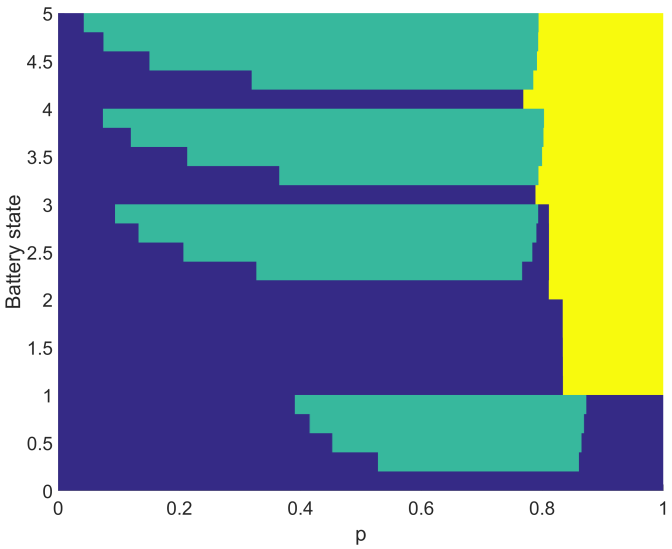


Fig. 2: Optimal thresholds for taking the actions D (blue), O (green), T (yellow) for $B_{max} = 5$, $\tau = 0.2$, $\beta = 0.98$, $\lambda_1 = 0.9$, $\lambda_0 = 0.6$, $R = 3$ and $q = 0.1$.

It is worth noticing the periodic-like behavior of the optimal policy observed in Fig. 2. One such behavior is the optimality of action D at the battery values which are whole numbers, followed by action O afterwards in mid range beliefs for those values of the battery exceeding 2. Value function corresponding to the parameters used in illustration of Fig. 2 is depicted in Fig. 3. Note that there is a significant increase in value function whenever the battery state is increased by one unit while value function almost does not change whenever the value of the battery state is confined between two consecutive whole number. Because of this, whenever the battery state of the transmitter is a whole number, by taking any other action other than deferring, there is a good chance that the state of the MDP makes a transition into a state with relatively lower value. Thus, it is better to choose action D in this case and not take the risk. However, when the battery state is between two consecutive whole number, it is safe to sense the channel, since at the worst case scenario when the channel is sensed to be in a bad state, the transmitter loses τ units but it makes a transition into a state which approximately have the same value. Thus at those values of the battery it is optimal to first sense the channel

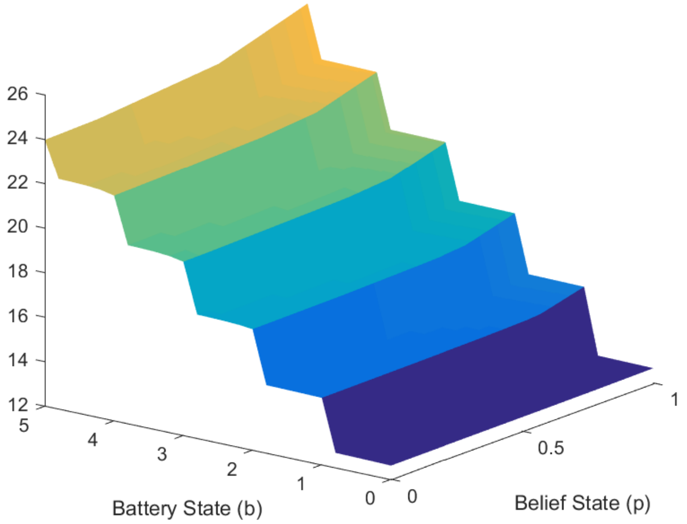
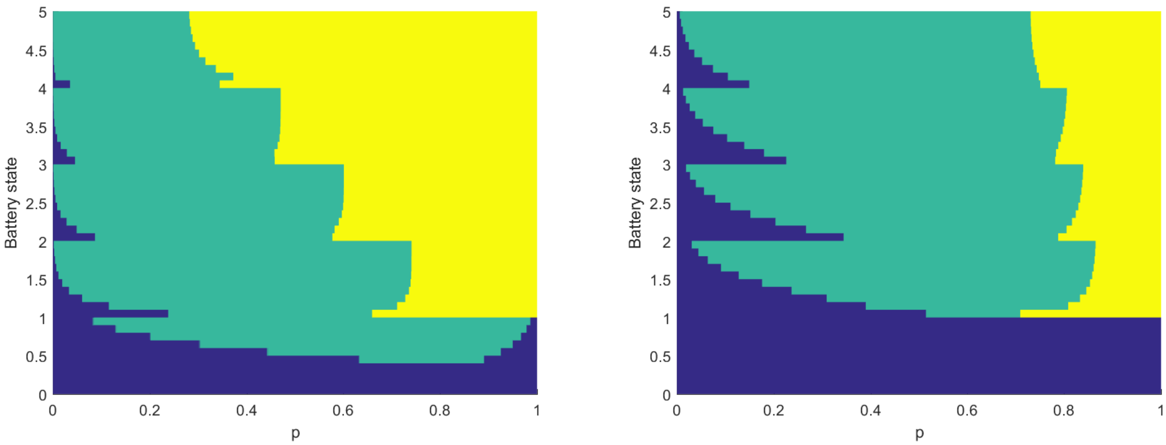


Fig. 3: Value function associated with $B_{max} = 5$, $\tau = 0.2$, $\beta = 0.98$, $\lambda_1 = 0.9$, $\lambda_0 = 0.6$, $R = 3$ and $q = 0.1$.

To investigate the effect of the EH rate, q, on the optimal transmission policy, consider the system parameters $B_{max} = 5$, $\tau = 0.1$, $\beta = 0.9$, $\lambda_1 = 0.8$, $\lambda_0 = 0.4$, and $R = 3$. We illustrate the optimal transmission policy for $q = 0.8$ and $q = 0.2$ in Fig. 4a and Fig. 4b, respectively. It can be observed by comparing those two figures that the yellow regions are much larger and blue areas are much more limited in Fig. 4a. This is because when the energy arrivals are more frequent,



(a) $q = 0.8$.

(b) $q = 0.2$.

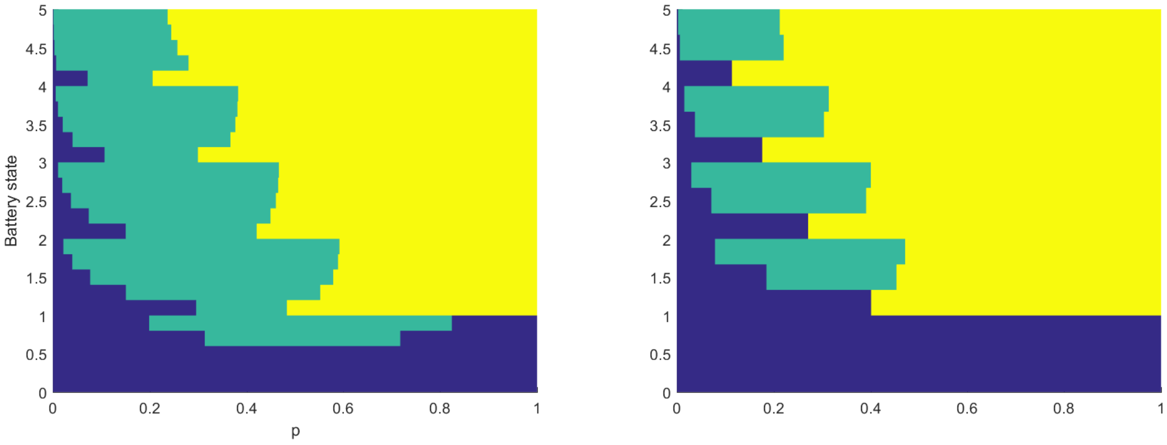
Fig. 4: Optimal thresholds for taking the actions D (blue), O (green), T (yellow) for $B_{max} = 5$, $\tau = 0.1$, $\beta = 0.9$, $\lambda_1 = 0.8$, $\lambda_0 = 0.4$, and $R = 3$.

the EH node tends to consume its energy more generously. We also observe that The EH node always defer its transmission for $b < 1$ when energy is limited (in Fig. 4b), whereas it opts for sensing the channel when energy is more abundant.

Next, we investigate the effect of the sensing cost, τ, on the optimal policy. To illustrate this effect, we choose the system parameters as $B_{max} = 5$, $\beta = 0.9$, $\lambda_1 = 0.8$, $\lambda_0 = 0.4$, $R = 3$ and $q = 0.8$. Optimal action regions are shown in Fig. 5a and Fig. 5b for sensing cost values $\tau = 0.2$ and $\tau = 1/3$, respectively. By comparing Fig. 5a and Fig. 5b, it is evident that a higher cost of sensing results in less incentive for sensing the channel. We observe that in Fig. 5b that the green areas has shrunk with respect to Fig. 5a, i.e, the transmitter is more likely to take a risk and transmit without sensing, or defer its defer its transmission when sensing consumes a significant portion of the available energy.

B. *Throughput performance*

In this section, we compare the performance of the optimal policy with two alternative policies, i.e., a greedy policy and a single-threshold policy. In the greedy policy, the EH node transmits whenever it has energy in its battery. In one threshold policy there are only two actions defer (D) or transmit (A). We optimize the threshold corresponding to each battery state for the single-threshold policy using the value iteration algorithm. By choosing the parameters $B_{max} = 5$, $\tau = 0.1$, $\beta = 0.999$, $\lambda_1 = 0.7$, $\lambda_0 = 0.2$, $R = 2$, the throughput achieved by these three policies



(a) $\tau = 0.2$.

(b) $\tau = 1/3$.

Fig. 5: Optimal thresholds for taking the actions D (blue), O (green), T (yellow) for $B_{max} = 5$, $\beta = 0.9$, $\lambda_1 = 0.8$, $\lambda_0 = 0.4$, $R = 3$ and $q = 0.8$.

are plotted in Fig. 6 with respect to the EH rate q. As expected we observe that the greedy policy performs the worst as it does not exploit the transmitter's knowledge about the state of the channel. We can see that by simply exploiting the ACK/NACK feedback from the receiver in order to defer the transmission, it is possible to achieve a higher throughput than the greedy policy at all values of the EH rate. On the other hand, by further introducing the channel sensing action the throughput of the system is substantially increased. The improvement is particularly higher for the mid-range of q values, for which the transmitter benefits more from the flexibility offered by three actions.

C. *Optimal policy evaluation with two different transmission rates*

Similar to Theorem 1, it is possible to show that for the case of two different transmission rates, the optimal policy is a threshold-type policy. However due to non-convexity of actions D and L it is not possible to elaborately characterize the optimal policy as in (14) and (15). Thus, we numerically evaluate the optimal policy in the following.

By assuming that $B_{max} = 5$, $\tau = 0.2$, $\beta = 0.95$, $\lambda_1 = 0.9$, $\lambda_0 = 0.6$, $R_1 = 2.1$, $R_2 = 3$ and $q = 0.5$, the optimal policy is illustrated in Fig. 7. In Fig. 7 each state $(b,\ p)$ is illustrated with a different color corresponding to the optimal policy at that state. In the figure, the areas highlighted with purple color correspond to those states at which deferring (D) the transmission is optimal, blue ares correspond to those states in which transmitting at low rate (L) is optimal,

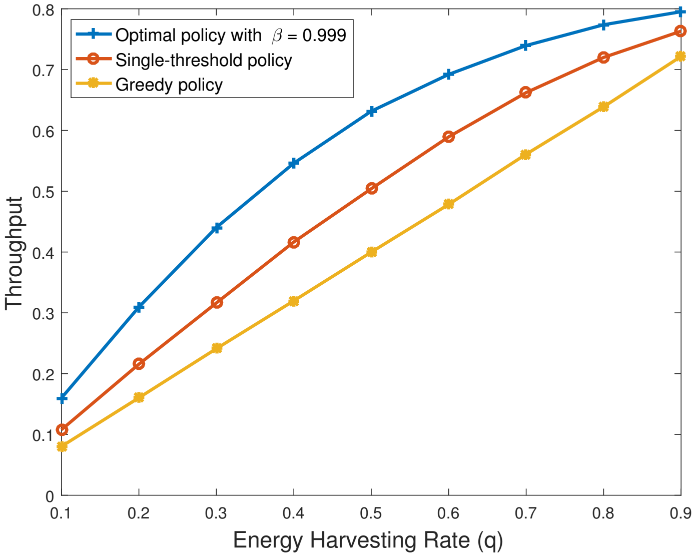


Fig. 6: Throughput comparison among the optimal, greedy, and single-threshold policies as a function of the EH rate, q.

green areas correspond to those states in which sensing and deferring is optimal (OD), turquoise areas correspond to the states at which sensing and transmitting opportunistically (OT) is optimal, and finally yellow areas correspond to the states for which transmitting without sensing (T) is optimal.

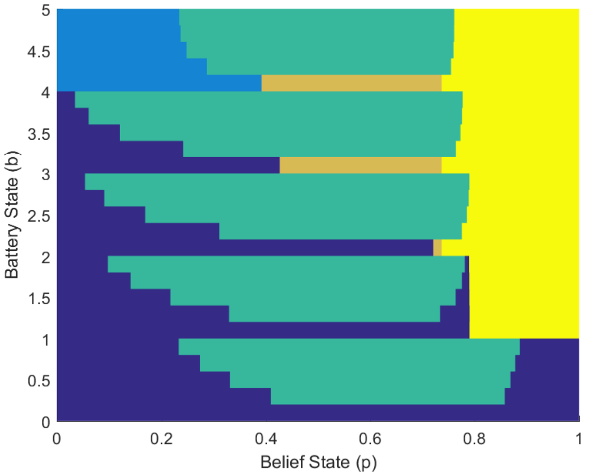


Fig. 7: Optimal thresholds for taking the actions D (purple), L (blue), OD (green), OT (turquoise), T (yellow) for $B_{max} = 5$, $\tau = 0.2$, $\beta = 0.95$, $\lambda_1 = 0.9$, $\lambda_0 = 0.6$, $R_1 = 2.1$, $R_2 = 3$ and $q = 0.5$.

The same periodic-like behavior observed in Fig. 2 also presents itself as in Fig. 7. As expected the optimal policy is again a battery dependent threshold-type policy with respect to the value of the belief. Although the optimal state space for action L (blue areas) is convex as seen in Fig. 7, in general this is not necessarily true.

VI. Conclusions and Future Work

In this work we considered an EH transmitter equipped with a battery, operating over a time varying finite-capacity wireless channel with memory, modeled as a Gilbert-Elliot channel. The transmitter receives ACK/NACK feedback after each transmission, which can be used to track the channel. We further consider channel sensing, which the transmitter can use to learn the current channel state a a certain energy and time cost. Therefore, at the beginning of each time slot, the transmitter has three possible actions to maximize the total expected discounted number of bits transmitted over an infinite time horizon: i) deferring the transmission to save its energy for future use, ii) transmitting at a rate of R bits, and iii) sensing the channel to reveal the current channel state by consuming a portion of its energy and time, followed by transmission at a reduced rate consuming the remainder of the energy unit, only if the channel is in the good state. We formulated the problem as a POMDP, which is then converted into a MDP with continuous state space by introducing a belief parameter for the channel state. Then we proved that the optimal policy is a threshold policy, where the threshold values as the belief parameter depends on the battery state. We find the optimal threshold values numerically using the value iteration policy. In terms of throughput, we compared the optimal policy to the alternative policies, the greedy policy and a single-threshold policy which does not have channel sensing capability. We have shown through simulations that the channel sensing capability improves the performance significantly, thanks to the increased adaptability to the channel conditions it provides. For future studies, we will consider the case where the sensing is not perfect. Another interesting problem is to consider the case in which the EH transmitter has the option to choose the duration of the sensing which determines its accuracy.

Appendix A

The Proof of Lemma 1

Proof. Define $V(b,p,n)$ as the optimal value function for the finite-horizon problem spanning only n time slots. We will first prove the convexity of $V(b,\ p,\ n)$ in p by induction. Optimal value function can be written as follows,

$$V(b,\ p,\ n) = \max\left\{V_D(b,\ p,n), V_O(b,\ p,n), V_A(b,\ p,n)\right\}, \tag{22}$$

where

$$\begin{aligned}
V_D(b,\ p,\ n) =& \beta\left[qV\left(b+1,\ J(p),\ n-1\right)+(1-q)V\left(b,\ J(p),\ n-1\right)\right], && (23)\\
V_O(b,\ p,\ n) =& p\left[(1-\tau)R+\beta\left(qV(b,\ \lambda_1,\ n-1)+(1-q)V(b-1,\ \lambda_1,\ n-1)\right)\right]\\
&+(1-p)\beta\left[V(\min\left\{b-\tau+1, B_{max}\right\},\ \lambda_0,\ n-1)+(1-q)V(b-\tau,\ \lambda_0,\ n-1)\right]\ b\geq 1,\\
& && (24)
\end{aligned}$$

$$\begin{aligned}
V_O(b,\ p,n) =& \beta\left[\,qpV(b-\tau+1,\ \lambda_1,\ n-1)\right.\\
&+q(1-p)V(b-\tau+1,\ \lambda_0,\ n-1)+(1-q)pV(b-\tau,\ \lambda_1,\ n-1)\\
&\left.+(1-q)(1-p)V(b-\tau,\ \lambda_0,\ n-1)\,\right],\quad \text{for } \tau\leq b<1, && (25)\\
V_A(b,\ p,\ n) =& p\left[R+\beta\left(qV(b,\ \lambda_1,\ n-1)+(1-q)V(b-1,\ \lambda_1,\ n-1)\right)\right]\\
&+(1-p)\beta\left[V(b,\ \lambda_0,\ n-1)+(1-q)V(b-1,\ \lambda_0,\ n-1)\right],\ b\geq 1. && (26)
\end{aligned}$$

Note that when $b<1$, we have $V(b,\ p,\ 1)=0$, and when $b\geq 1$ we have $V(b,\ p,\ 1)=\max\left\{pR,\ (1-\tau)pR, 0\right\}=pR$. We see that $V(b,\ p,\ 1)$ is a convex function of p.

Now, let us assume that $V(b,\ p,\ n-1)$ is convex in p for any $b\geq 0$, then for $a\in[0,\ 1]$ we can investigate the convexity of the value function for each action separately as follows.

For deferring the transmission, i.e., $T=D$, we can write:

$$\begin{aligned}
V_D\left(b,\ ap_1+(1-a)p_2,n\right) =& \beta\left[\,qV\left(b+1,\ J(ap_1+(1-a)p_2),n-1\right)\right.\\
&\left.+(1-q)V\left(b,\ J(ap_1+(1-a)p_2),\ n-1\right)\,\right]\\
=& \beta\left[\,qV\left(b+1,\ aJ(p_1)+(1-a)J(p_2),n-1\right)\right.\\
&\left.+(1-q)V\left(b,\ aJ(p_1)+(1-a)J(p_2),n-1\right)\,\right]\\
\leq& a\beta\left[\,qV\left(b+1,\ J(p_1),n-1\right)\right.\\
&\left.+(1-q)V\left(b,\ J(p_1),n-1\right)\,\right]\\
&+(1-a)\beta\left[\,qV\left(b+1,\ J(p_2),n-1\right)\right.\\
&\left.+(1-q)V\left(b,\ J(p_2),n-1\right)\,\right]
\end{aligned}$$

$$
\begin{aligned}
&= aV_D(b,\ p_1,\ n) + (1-a)V_D(b,\ p_2,\ n)\\
&\leq aV(b,\ p_1,\ n) + (1-a)V(b,\ p_2,\ n) \qquad (27)
\end{aligned}
$$

Hence, $V_D(b,\ p,\ n)$ is convex in p. Note that $V_O(b,\ p,\ n)$ and $V_A(b,\ p,\ n)$ are linear functions of p, thus they are obviously convex in p.

Since the value function $V(b,\ p,\ n)$ is the maximum of three (or, in some cases two) of the convex functions when $b \geq 1$ ($\tau \leq b < 1$), it is also convex.

By induction we can claim the convexity of $V(b,\ p,\ n)$ for all n. Since $V(b,\ p,\ n) \to V(b,\ p)$ as $n \to \infty$, $V(b,\ p)$ is also convex. □

APPENDIX B
THE PROOF OF LEMMA 2

Proof. We will again use induction to prove the claim for $V(b,\ p,\ n)$ defined as in Appendix A as the optimal value function when the decision horizon spans n stages. We have $V(b,p,1) = 0$ if $b < 1$ and we have $V(b,p,1) = pR$ if $b \geq 1$,. Hence, $V(b,p,1)$ is trivially non-decreasing in b. Suppose that $V(b,p,n-1)$ is non-decreasing in b. Each of the value functions given in (23), (24), (25) and (26) is the summation of positive weighted non-decreasing functions. Therefore, they are all non-decreasing in b. Since the optimal value function is the maximum of this non-decreasing functions, it is also non-decreasing in b for any n. Similarly to Appendix A, by letting $n \to \infty$ to infinity we conclude that $V(b,\ p)$ is non-decreasing in b. □

APPENDIX C
THE PROOF OF LEMMA 3

Proof. We employ induction on $V(b,\ p,\ n)$ once again. For $n = 1$, $V(b,\ p,\ 1)$ is 0 if $b < 1$, and pR if $b \geq 1$. Therefore, $V(b,\ p,\ 1)$ is non-decreasing in p for any given b.

Assume that $V(b,\ p,\ n-1)$ is non-decreasing in p. Since $J(p)$ is non-decreasing, it is easy to see that $V_D(b,\ p,\ n)$ in (23) is also non-decreasing.

Since $V_O(b,\ p,\ n)$ and $V_A(b,\ p,\ n)$ are linear in p, we have $V_A(b,\ ap_1 + (1-a)p_0) = aV_A(p_1) + (1-a)V_A(p_0)$ and $V_O(b,\ ap_1 + (1-a)p_0) = aV_O(p_1) + (1-a)V_O(p_0)$. Using this result, we have

$$
\begin{aligned}
V_A(b,\ p_1,n) - V_A(b,\ p_0,n) &= V_A(b,\ p_1 - p_0 + p_0,n) - V_A(b,\ p_0,n) \qquad (28a)\\
&= V_A(b,\ p_1 - p_0,n) \geq 0. \qquad (28b)
\end{aligned}
$$

Note that $(28b)$ follows from the fact that $V_A(b,\ p_1-p_0+p_0,n)=V_A(b,\ p_1-p_0,n)+V_A(b,\ p_0,n)$. Similarly for transmitting opportunistically we have

$$\begin{aligned} V_O(b,\ p_1,n)-V_O(b,\ p_0,n) =& V_O(b,\ p_1-p_0+p_0,n)-V_O(b,\ p_0,n) \\ =& V_O(b,\ p_1-p_0,n)\geq 0. \end{aligned} \tag{29}$$

Since the value function, $V(b,\ p,\ n)$, is the maximum of three non-decreasing functions, it is also non-decreasing. Hence, by letting $n \rightarrow \infty$, we prove that $V(b,p)$ is non-decreasing in p. □

References

[1] J.A. Paradiso and T. Starner. Energy scavenging for mobile and wireless electronics. *IEEE Pervasive Computing*, 4(1):18–27, Jan. 2005.

[2] G. Park, T.Rosing, M.D. Todd, C.R. Farrar, and W. Hodgkiss. Energy harvesting for structural health monitoring sensor networks. *Journal of Infrastructure Systems*, 14(1):64–79, Mar. 2008.

[3] D. Gunduz, K. Stamatiou, N. Michelusi, and M. Zorzi. Designing intelligent energy harvesting communication systems. *IEEE Communications Magazine*, 52(1):210–216, Jan. 2014.

[4] H. Li, C. Huang, P. Zhang, S. Cui, and J. Zhang. Distributed opportunistic scheduling for energy harvesting based wireless networks: A two-stage probing approach. *IEEE/ACM Transactions on Networking*, 24(3):1618–1631, Jun. 2016.

[5] Q. Zhang and S. A. Kassam. Finite-state Markov model for Rayleigh fading channels. *IEEE Trans. on Communs*, 47(11):1688–1692, Nov. 1999.

[6] E. N. Gilbert. Capacity of a burst-noise channel. *The Bell System Technical Journal*, 39(5):1253–1265, Sep. 1960.

[7] M. L. Ku, Y. Chen, and K. J. R. Liu. Data-driven stochastic models and policies for energy harvesting sensor communications. *IEEE Journal on Selected Areas in Communications*, 33(8):1505–1520, Aug. 2015.

[8] N. Michelusi, K. Stamatiou, and M. Zorzi. Transmission policies for energy harvesting sensors with time-correlated energy supply. *IEEE Transactions on Communications*, 61(7):2988–3001, Jul. 2013.

[9] A. Hentati, F. Abdelkefi, and W. Ajib. Energy allocation for sensing and transmission in WSNs with energy harvesting Tx/Rx. In *IEEE Vehicular Technology Conf. (VTC Fall),*, pages 1–5, Sep. 2015.

[10] S. Mao, M. H. Cheung, and V. W. S. Wong. Joint energy allocation for sensing and transmission in rechargeable wireless sensor networks. *IEEE Transactions on Vehicular Technology*, 63(6):2862–2875, Jul. 2014.

[11] P. Blasco and D. Gunduz. Multi-access communications with energy harvesting: A multi-armed bandit model and the optimality of the myopic policy. *IEEE Journal on Selected Areas in Communications*, 33(3):585–597, Mar. 2015.

[12] C. K. Ho and R. Zhang. Optimal energy allocation for wireless communications with energy harvesting constraints. *IEEE Transactions on Signal Processing*, 60(9):4808–4818, Sep. 2012.

[13] B. T. Bacinoglu and E. Uysal-Biyikoglu. Finite-horizon online transmission scheduling on an energy harvesting communication link with a discrete set of rates. *Journal of Communications and Networks*, 16(3):393–300, Jun. 2014.

[14] M. S. H. Abad, D. Gunduz, and O. Ercetin. Energy harvesting wireless networks with correlated energy sources. In *2016 IEEE Wireless Communications and Networking Conference*, pages 1–6, Apr. 2016.

[15] M. Kashef and A. Ephremides. Optimal packet scheduling for energy harvesting sources on time varying wireless channels. *Journal of Communications and Networks*, 14(2):121–129, Apr. 2012.

[16] A. Aprem, C. R. Murthy, and N. B. Mehta. Transmit power control policies for energy harvesting sensors with retransmissions. *IEEE Journal of Selected Topics in Signal Processing*, 7(5):895–906, Oct. 2013.

[17] Q. Zhao, L. Tong, A. Swami, and Y. Chen. Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework. *IEEE Journal on Selected Areas in Communications*, 25(3):589–600, Apr. 2007.

[18] A. Laourine and L. Tong. Betting on Gilbert-Elliot channels. *IEEE Transactions on Wireless Communications*, 9(2):723–733, Feb. 2010.

[19] William S. Lovejoy. A survey of algorithmic methods for partially observed Markov decision processes. *Annals of Operations Research*, 28(1):47–65, Dec. 1991.

[20] P. Blasco, D. Gunduz, and M. Dohler. A learning theoretic approach to energy harvesting communication system optimization. *IEEE Transactions on Wireless Communications*, 12(4):1872–1882, Apr. 2013.

[21] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994.