Optimal Radio Frequency Energy Harvesting with Limited Energy Arrival Knowledge

Zhenhua Zou, Anders Gidmark, Themistoklis Charalambous and Mikael Johansson

***Abstract*—We develop optimal sleeping and harvesting policies for radio frequency (RF) energy harvesting (EH) devices, formalizing the following intuition: when the ambient RF energy is low, devices consume more energy being awake than what can be harvested, and should enter sleep mode; when the ambient RF energy is high, on the other hand, it is essential to wake up and harvest. Towards this end, we consider a scenario with intermittent energy arrivals described by a two-state Gilbert-Elliott Markov chain model. The challenge is that the state of the Markov chain can only be observed during the harvesting action, and not while in sleep mode. Two scenarios are studied under this model. In the first scenario, we assume that the transition probabilities of the Markov chain are known and formulate the problem as a Partially Observable Markov Decision Process (POMDP). We prove that the optimal policy has a threshold structure and derive the optimal decision parameters. In the practical scenario where the ratio between the reward and the penalty is neither too large nor too small, the POMDP framework and the threshold-based optimal policies are very useful for finding non-trivial optimal sleeping times. In the second scenario, we assume that the Markov chain parameters are unknown and formulate the problem as a Bayesian adaptive POMDP, and propose a heuristic posterior sampling algorithm to reduce the computational complexity. The performance of our approaches is demonstrated via numerical examples.**

***Index Terms*—Energy harvesting, ambient radio frequency energy, Partially Observable Markov Decision Process, Bayesian inference, learning.**

I. Introduction

One important concept in green communications is to use renewable energy sources to replenish the batteries of network nodes and exploit ambient energy as an alternative power source. Radio frequency (RF) energy harvesting (EH) is one of the energy harvesting methods that have recently attracted a lot of attention (see, for example, [1]–[3] and references therein). In RF-EH, a device can capture ambient RF radiation from a variety of radio transmitters (such as television/radio broadcast stations, WiFi, cellular base stations and mobile phones) and convert it into a direct current through rectennas [4], see Figure 1. It has been shown that low-power wireless systems such as wireless sensor networks with RF energy harvesting capabilities can have a significantly prolonged lifetime, even to the point where they can become self-sustained and support previously infeasible applications [5].

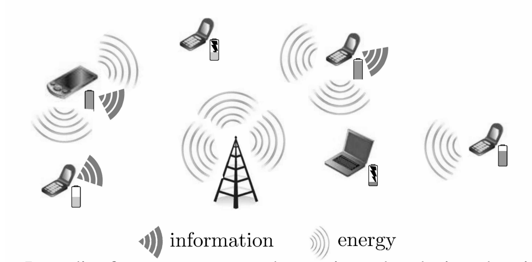

Fig. 1. In radio frequency energy harvesting, the device that is not the destination of the packet can capture RF radiation of the wireless transmission from cellular communication, WiFi or TV towers, and convert it into a direct current through rectennas.

However, in many cases the RF energy is intermittent. This can be due to temporary inactive periods of communication systems with bursty traffic and/or due to multi-path fading in the wireless channel [6]. Moreover, the energy spent by wireless devices to wake up the radio and assess the current channel condition is non-negligible. Hence, when the ambient energy is low, staying awake and harvesting energy can result in an energy loss, and it is better to sleep. The challenge in the energy harvesting process lies in the fact that the wireless device does not know the energy level before it attempts to harvest. For this reason, it is crucial to develop policies for deciding if a wireless node should harvest or sleep to maximize the accumulated energy.

In this paper, we study the problem of energy harvesting for a single wireless device in an environment where the ambient RF energy is intermittent. Energy harvesting with intermittent energy arrivals has recently been investigated under the assumption that the energy arrivals are described by *known* Markov processes [7]–[11]. However, in practice the energy arrivals may not follow the chosen Markov process model. It is therefore necessary that harvesting policies do not presume the arrival model, but allow for unknown energy arrivals. Towards this direction, the optimal harvesting problem has only been targeted via the classical Q-learning method in [12]. The Robbins-Monro algorithm, the mathematical cornerstone of Q-learning, was applied in [13] to derive optimal policies with a faster convergence speed by exploiting that the optimal policy is threshold-based. However, both the Q-learning method and the Robbins-Monro algorithm rely on heuristics (e.g., ϵ-greedy) to handle the exploration-exploitation trade-off [14]. The optimal choice of the step-size that attains the best convergence speed is also not clear; only a set of sufficient conditions for asymptotic convergence is given in the literature.

This work was performed when Z. Zou was with the Automatic Control Lab, School of Electrical Engineering, Royal Institute of Technology (KTH), Stockholm, Sweden in 2014 and the Department of Signals and Systems, Chalmers University of Technology, Gothenburg, Sweden in 2015. (Email: zzou@kth.se).

T. Charalambous is with the Department of Signals and Systems, Chalmers University of Technology, Gothenburg, Sweden (Email: thecha@chalmers.se).

A. Gidmark and M. Johansson are with the Automatic Control Lab, School of Electrical Engineering, Royal Institute of Technology (KTH), Stockholm, Sweden (Emails: {gidmark,mikaelj}@kth.se).

All the aforementioned works assume that the energy arrival state is known at the decision maker *before* the decision is taken. This is an unrealistic assumption, since it does not take into account the energy cost for the node to wake up and track the energy arrival state, and being active continuously can be detrimental when the ambient energy level is low. Partial observability issues in energy harvesting problems have only been considered in scenarios related to knowledge about battery State-of-Charge [15], event occurrences in optimal sensing problems [16], and channel state information for packet transmissions [17]. To the best of our knowledge, neither the scenario with partial observability of the energy arrival nor this scenario combined with an unknown model have been addressed in the literature before.

Due to the limited energy arrival knowledge and the cost for unsuccessful harvesting, the fundamental question being raised is whether and when it is beneficial for a wireless device to try and harvest energy from ambient energy sources. In this paper, we aim at answering this question by developing optimal sleeping and harvesting policies that maximize the accumulated energy. More specifically, we make the following contributions:

- We model the energy arrivals using an abstract two-state Markov chain model where the node receives a reward in the good state and incurs a cost in the bad state. The state of the model is revealed to the node only if it chooses to harvest. In absence of new observations, future energy states are predicted based on knowledge about the transition probabilities of the Markov chain.
- We propose a simple yet practical reward function that encompasses the effects of the decisions made based on the states of the Markov chain.
- We study the optimal energy harvesting problem under two assumptions on the parameters of the energy arrival model.
  1) For the scenario where the parameters are known, we formulate the problem of whether to harvest or to sleep as a Partially Observable Markov Decision Process (POMDP). We show that the optimal policy has a threshold structure: after an unsuccessful harvesting, the optimal action is to sleep for a constant number of time-slots that depends on the parameters of the Markov chain; otherwise, it is always optimal to harvest. The threshold structure leads to an efficient computation of the optimal policy. Only a handful of papers have explicitly characterized the optimality of threshold-based policies for POMDP (for example, [18], [19]) and they do not deal with the problem considered in this work.
  2) For the scenario when the transition probabilities of the Markov chain are not known, we apply a novel Bayesian online-learning method. To reduce the complexity of the computations, we propose a heuristic posterior sampling algorithm. The main idea of Bayesian online learning is to specify a prior distribution over the unknown model parameters, and update a posterior distribution by Bayesian inference over these parameters to incorporate new information about the model as we choose actions and observe results. The exploration-exploitation dilemma is handled directly as an explicit decision problem modeled by an extended POMDP, where we aim to maximize future expected utility with respect to the current uncertainty on the model. The other advantage is that we can define an informative prior to incorporate previous beliefs about the parameters, which can be obtained from, for example, domain knowledge and field tests. Our work is the first in the literature that introduces and applies the Bayesian adaptive POMDP framework [20] in energy harvesting problems with unknown state transition probabilities.
- The schemes proposed in this paper are evaluated in simulations and significant improvements are demonstrated compared to letting nodes harvest continuously or at random times.

We would like to stress that the approach we develop in this paper is not limited to RF-EH, but can be extended to other forms of renewable resources, such as solar and piezoelectric. The rest of this paper is organized as follows. The system model and the energy harvesting problem are introduced in Section II. In Section III we address the case of *known* Markov chain parameters, and derive the optimal sleeping and harvesting policies using POMDP; the threshold-based structure of the optimal policies are also shown. In Section IV we address the case of *unknown* Markov chain parameters and propose a Bayesian online learning method. Numerical examples are provided in Section V. Finally, in Section VI we draw conclusions and outline possible future research directions. All the proofs are presented in the Appendix.

II. System Model

We consider a single wireless device with the capability of harvesting energy from ambient energy sources. We assume that the overall energy level is constant during one time-slot, and may change in the next time-slot according to a two-state Gilbert-Elliott Markov chain model [21], [22]; see Fig. 2. In

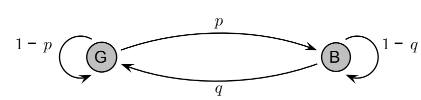

Fig. 2. Two-state Gilbert-Elliott Markov chain model.

this model, the good state (G) denotes the presence of energy to be harvested and the bad state (B) denotes the absence of energy to be harvested. The transition probability from the G state to B state is p, and the transition probability from B state to G state is q. The probabilities of staying at states G and B are $1-p$ and $1-q$, respectively. It can be easily shown that the steady state distribution of the Markov chain at B and G states are $p/(p+q)$ and $q/(p+q)$, respectively.

We consider the model in which the probability of being in state G in the next time-slot is higher if the current state is G, and the probability of being in state B in the next time-slot is higher if the current state is B. This requires that $1-p>q$ and corresponds to that RF energy is positively correlated in time, which is a valid assumption in practice.

At each time-slot, the node can take one of two possible actions: to harvest or to sleep. If the node chooses to harvest and the Markov chain is in the G state, a reward $r_1 > 0$ is received that represents the energy successfully harvested. If the Markov chain is in the B state during the harvesting action, a penalty $-r_0 < 0$ is incurred that represents the energy cost required to wake up the radio and try to detect if there exists any ambient energy to harvest. On the other hand, if the node sleeps, no reward is received. Therefore, the reward function is defined as

$$R(s,a) \triangleq \begin{cases} r_1, & (a=\mathcal{H}) \wedge (s=G), \\ -r_0, & (a=\mathcal{H}) \wedge (s=B), \\ 0, & a=\mathcal{S}, \end{cases} \tag{1}$$

where a denotes the action to harvest ($\mathcal{H}$) or sleep ($\mathcal{S}$), and s is the current Markov chain state.

Remark 1. *Note that one could impose a cost for sleeping. However, this does not change the problem setup since we could normalize the rewards and costs so that the sleeping cost is zero.*

Remark 2. *In addition, the choice of the exact numbers for r_0 and r_1 depend on hardware specifications, such as the energy harvesting efficiency and the energy harvesting cost. Note that r_0 represents a penalty incurred due to additional energy spent to turn on the RF-EH circuit. For larger static devices, one can possibly use large and specially tailored rectennas to harvest passively, so r_0 would be essentially zero; for small mobile devices, however, one is often bound to use a simple antenna combined with a power management circuit and ESD. The harvested signals are in general of very low power. Since the signals need to be rectified, a large percentage of the signal power is lost in the rectification bridge. If very little energy is harvested, one may get less energy from the antenna than what is needed to perform the DC/DC conversion, i.e., an energy penalty is incurred. These arguments and more details can be found in several papers; see, for example, [23, § IV.D] and [24, § III].*

Remark 3. *Even though in reality the energy harvested and hence the reward r_1 is not fixed, the choice of r_1 can be seen as the minimum; similarly, r_0 can be seen as the maximum. If we further assume that the successfully harvested and lost energy levels are i.i.d. random variables, then we can let r_1 and r_0 be the average harvested or lost energy, respectively.*

The state information of the underlying Markov chain can only be observed under the harvesting action, but it is inefficient to always harvest since there is a cost associated with an unsuccessful energy harvesting. On the other hand, the sleeping action does not incur any cost (apart from missing out a chance to harvest), but it does not reveal the state information. Thus, it is not immediately clear when it is better to harvest to maximize the reward. Furthermore, the transition probabilities of the Markov chain may not be known a priori, which makes the problem of maximizing the reward even more challenging.

Let $a_t \in \{\mathcal{H}, \mathcal{S}\}$ denote the action at time t, s_t denote the state of the Markov chain at time t, and $z_t \in \{G, B, Z\}$ denote the observation at time t where Z means no observation of the Markov chain. Let $a^t \triangleq \{a_0, a_1, \ldots, a_t\}$ denote the history of actions and $z^t \triangleq \{z_0, z_1, \ldots, z_t\}$ denote the history of observations. A policy π is a function that prescribes an action at time t based on the history of actions and observations up to time $t-1$. The goal is then to find the optimal policy $\pi^\star$ that maximizes the expected total discounted reward,

$$\pi^\star \in \arg\max_{\pi} \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R_t(s_t, a_t) \right],$$

where R_t is the reward at time t and the expectation is taken with respect to the transitions of the Markov chain. The discount factor $\gamma \in [0,1)$ models the importance of the energy arrivals at different time-slots in which the energy harvested in the future will be discounted. The discount factor can also be seen as a scenario where the node terminates its operation at each time-slot independently with probability $(1-\gamma)$ [25].

III. Optimal structured policy with unknown Markovian states

In this section, we first derive the optimal policy under *known* transition probabilities but *unknown* Markovian states by formulating it as a Partially Observable Markov Decision Process (POMDP) [26]. We show that the optimal policy has a threshold-based structure. This structural result simplifies both the off-line computations during the design phase and the real-time implementation.

A. POMDP formulation

Although the exact state is not known at each time-slot, we can keep a probability distribution (i.e., a belief) of the state based on the past observations and the knowledge on the Markov chain. We let b denote the belief, i.e., the probability that the RF energy state is good at the *current* time-slot, and b' is the belief at the *next* time-slot. In this way, at any given time the probability that the RF energy state is good is b and the probability that the state is bad is $1-b$. If we decide to harvest, then we will be able to observe the current RF energy state. If this state is good, then $b' = 1-p$ due to the definition of the Markov chain transition probabilities. If this state is bad, then $b' = q$ by the same reasoning. In the sleep state, we do not get any new information and the belief is updated using the Markov chain model parameters, i.e.,

$$b' = b(1-p) + (1-b)q = q + (1-p-q)b, \tag{2}$$

which is the probability of being at good state at the next time-slot given the probability at the current time-slot. This update converges to the stationary distribution of the good state. In summary, we have the following state transition probability

$$\mathbb{P}(b'|a=\mathcal{H}, b) = \begin{cases} b & \text{if } b' = 1-p, \\ 1-b & \text{if } b' = q; \end{cases} \tag{3}$$

$$\mathbb{P}(b'|a=\mathcal{S}, b) = \begin{cases} 1 & \text{if } b' = q + (1-p-q)b, \\ 0 & \text{otherwise}. \end{cases} \tag{4}$$

Note, the belief b takes values between q and $1-p$ since $1-p > q$ by assumption.

By Equation (1), the expected reward with belief b is

$$R(b,a) = bR(1,a) + (1-b)R(0,a) = \begin{cases} (r_0+r_1)b - r_0, & a = \mathcal{H}, \\ 0, & a = \mathcal{S}. \end{cases} \tag{5}$$

It has been shown in [26] that the belief b is a sufficient statistic for decision making given the past action history a^t and the past observation history z^t, and we can convert the POMDP to a corresponding MDP with the belief as the state. Hence, the policy π is also a function that prescribes an action a for the belief b. The expected total discounted reward for a policy π starting from initial belief b_0, also termed as the value function, is then

$$V^\pi(b_0) \triangleq \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t R_t(b_t,a_t)|b_0].$$

Since the state space is countable and the action space is finite with only two actions, there exists an optimal deterministic stationary policy $\pi^\star$ for any b [25, Theorem 6.2.10] such that

$$\pi^\star \in \arg\max_\pi V^\pi(b).$$

B. Optimal policy - value iteration

Let $V^\star \triangleq V^{\pi^\star}$ be the optimal value function. The optimal policy can be derived from the optimal value function, i.e., for any b, we have

$$\pi^\star(b) \in \arg\max_{a\in\{\mathcal{H},\mathcal{S}\}} \left[R(b,a) + \gamma \sum_{b'} \mathbb{P}(b'|a,b) V^\star(b') \right].$$

The problem of deriving the optimal policy is then to compute the optimal value function. It is known that the optimal value function satisfies the Bellman optimality equation [25, Theorem 6.2.5],

$$V^\star(b) = \max_{a\in\{\mathcal{H},\mathcal{S}\}} \left[R(b,a) + \gamma \sum_{b'} \mathbb{P}(b'|a,b) V^\star(b') \right],$$

and the optimal value function can be found by the value iteration method shown in Algorithm 1. The algorithm utilizes the fixed-point iteration method to solve the Bellman optimality equation with stopping criteria. If we let $t \to \infty$, then the algorithm returns the optimal value function $V^\star(b)$ [25].

Algorithm 1: Value iteration algorithm [25]

Input: Error bound ϵ
Output: $V(b)$ with $\sup_b |V(b) - V^\star(b)| \le \epsilon/2$.

```
Initialization: At t = 0, let V_0(b) = 0 for all b
repeat
    Compute V_{t+1}(b) for all states b,
        V_{t+1}(b) = max_{a∈{H,S}} [ R(b,a) + γ Σ_{b'} P(b'|a,b) V_t(b') ].
    Update t = t + 1.
until sup_b |V_{t+1}(b) − V_t(b)| ≤ ε(1 − γ)/2γ.
```

C. Optimality of the threshold-based policy

Let $Q_{t+1}(b,a)$ denote the value function of any action $a \in \{\mathcal{H},\mathcal{S}\}$ in Algorithm 1,

$$Q_{t+1}(b,a) \triangleq R(b,a) + \gamma \sum_{b'} \mathbb{P}(b'|a,b) V_t(b'), \tag{6}$$

and let $Q_\infty(b,a) = \lim_{t\to\infty} Q_t(b,a)$. We first show that the optimal policy has a threshold structure:

Proposition 1. *Define*

$$\bar{b} \triangleq \min_b \{Q_\infty(b,\mathcal{H}) \ge Q_\infty(b,\mathcal{S})\}. \tag{7}$$

If $\bar{b} < q/(p+q)$, then the optimal policy is to continue to harvest after a successful harvesting time slot, and to sleep for

$$N \triangleq \left\lceil \log_{1-p-q} \frac{q-(p+q)\bar{b}}{q} \right\rceil - 1 \tag{8}$$

time-slots after an unsuccessful harvesting. If the threshold $q/(p+q) \le \bar{b} \le 1-p$, then the optimal policy is to continue to harvest after a successful harvesting, and to never harvest following an unsuccessful harvesting. If the threshold $1-p < \bar{b}$, then the optimal policy is to not harvest at all.

Remark: The sleeping time N is the minimum time after which the belief exceeds the threshold $\bar{b}$ after an unsuccessful harvesting attempt. Details can be found in the proof in the appendix.

Computing the optimal threshold $\bar{b}$ as suggested by Proposition 1 would be computationally demanding. However, by leveraging on the fact that the optimal policy has a threshold structure, we will, in Proposition 2, show that it is possible to derive a computationally efficient expression for the optimal waiting time N based on channel statistics and rewards.

Note that if $q/(p+q) \le \bar{b} \le 1-p$, then the optimal policy will eventually stop harvesting since there is a non-zero probability of an unsuccessful harvesting attempt. Hence, in Proposition 2, we focus on the more interesting case where $\bar{b} < q/(p+q)$.

Proposition 2. *Suppose $\bar{b} < q/(p+q)$. Let $b' \triangleq q[1-(1-p-q)^{n+1}]/(p+q)$, let $F(n) \triangleq \gamma^{n+1} r_1(b'-1+p) + r_1 - p(r_0+r_1)$, and let $G(n) \triangleq \gamma^{n+1}(b'(1-\gamma) - (1-\gamma+\gamma p)) + 1 - \gamma + \gamma p$. The optimal policy is to continuously harvest after a successful harvesting, and to sleep for*

$$N \triangleq \arg\max_{n\in\{0,1,\dots\}} \frac{F(n)}{G(n)} \tag{9}$$

time-slots after an unsuccessful harvesting attempt.

IV. Bayesian online learning unknown transition probabilities

In many practical scenarios, the transition probabilities of the Markov chain that models the energy arrivals may be initially unknown. To obtain an accurate estimation, we need to sample the channel many times, a process which unfortunately consumes a large amount of energy and takes a lot of time. Thus, it becomes crucial to design algorithms that balance the parameter estimation and the overall harvested energy; this is the so-called exploration and exploitation dilemma. Towards this end, in this section, we first formulate the optimal energy harvesting problem with unknown transition probabilities as a Bayesian adaptive POMDP [20]. Next, we propose a heuristic posterior sampling algorithm based on the threshold structure of the optimal policy with known transition probabilities. The Bayesian approach can incorporate the domain knowledge by specifying a proper prior distribution of the unknown parameters. It can also strike a natural trade-off between exploration and exploitation during the learning phase.

A. Models and Bayesian update

The Beta distribution is a family of distributions that is defined on the interval $[0,1]$ and parameterized by two parameters. It is typically used as conjugate prior distributions for Bernoulli distributions so that the posterior update after observing state transitions is easy to compute. Hence, for this work, we assume that the unknown transition probabilities p and q have independent prior distributions following the Beta distribution parameterized by $\phi \triangleq [\phi_1 \ \phi_2 \ \phi_3 \ \phi_4]^T \in \mathbb{Z}_+^4$:

$$\begin{aligned}\mathbb{P}(p,q;\phi) &= \mathbb{P}(p,q;\phi_1,\phi_2,\phi_3,\phi_4)\\ &\overset{(a)}{=} \mathbb{P}(p;\phi_1,\phi_2)\mathbb{P}(q;\phi_3,\phi_4),\end{aligned} \tag{10}$$

where (a) stems from the fact that p and q have independent prior distributions. The Beta densities of probabilities p and q are given by

$$\mathbb{P}(p;\phi_1,\phi_2) = \frac{\Gamma(\phi_1+\phi_2)}{\Gamma(\phi_1)\Gamma(\phi_2)} p^{\phi_1-1}(1-p)^{\phi_2-1}, \tag{11}$$

$$\mathbb{P}(q;\phi_3,\phi_4) = \frac{\Gamma(\phi_3+\phi_4)}{\Gamma(\phi_3)\Gamma(\phi_4)} q^{\phi_3-1}(1-q)^{\phi_4-1}, \tag{12}$$

respectively, where $\Gamma(\cdot)$ is the gamma function, given by $\Gamma(y) = \int_0^\infty x^{y-1}e^{-x}dx$. However, for $y \in \mathbb{Z}_+$ (as it is the case in our work), the gamma function becomes $\Gamma(y) = (y-1)!$.

By using the Beta distribution parameterized by posterior counts for p and q, the posterior update after observing state transitions is easy to compute. For example, suppose the posterior count for the parameter p is $\phi_1 = 5$ and $\phi_2 = 7$. After observing state transitions from G to B (with probability p) for 2 times and state transitions from G to G (with probability $1-p$) for 3 times, the posterior count for the parameter p is simply $\phi_1 = 5+2 = 7$ and $\phi_2 = 7+3 = 10$. Without loss of generality, we assume that ϕ initially is set to $[1,1,1,1]$ to denote that the parameters p and q are between zero and one with equal probabilities, i.e., $\mathbb{P}(p,q) = 1$.

Note that we can infer the action history a^t from the observation history z^t. More specifically, for each time t, if $z_t = Z$, then $a_t = \mathcal{S}$, and if $z_t \in \{G,B\}$, then $a_t = \mathcal{H}$. In what follows, we use only the observation history z^t for posterior update for the sake of simplicity. Consider the joint posterior distribution $\mathbb{P}(s_t,p,q|z^{t-1})$ of the energy state s_t and the transition probability p and q at time t from the observation history z^{t-1}. Let

$$S(z^{t-1}) = \{s^{t-1} : s_\tau = z_\tau \ \forall \tau \in \{t' : z_{t'} \neq Z\}\} \tag{13}$$

denote all possible state histories based on the observation history z^{t-1}.

For a given $s^{t-1} \in S(z^{t-1})$ and $s^t = \{s^{t-1}, s_t\}$, suppose there are in total a number transitions from G to B, b number of transitions from G to G, c number of transitions from B to G, and d number of transitions from B to B. We have that

$$\mathbb{P}(s^t|p,q)\mathbb{P}(p,q) = \mathbb{P}(s^t|p,q) = p^a(1-p)^b q^c (1-q)^d.$$

Since, multiple state histories can lead to the same number of transitions, we let $C(\phi, S(z^{t-1}), s_t)$ denote the total number of state histories that lead to the posterior count ϕ from the initial condition that all counts are equal one, and we call it the *appearance count* to distinguish it from the posterior count ϕ. Hence,

$$\begin{aligned}&\mathbb{P}(s_t,p,q|z^{t-1})\mathbb{P}(z^{t-1})\\ &= \mathbb{P}(z^{t-1},s_t|p,q)\mathbb{P}(p,q) = \sum_{s^{t-1}} \mathbb{P}(z^{t-1},s^t|p,q)\mathbb{P}(p,q)\\ &= \sum_{s^{t-1}\in S(z^{t-1})} \mathbb{P}(s^t|p,q)\mathbb{P}(p,q)\\ &= \sum_{\phi} C(\phi,S(z^{t-1}),s_t) p^{\phi_1-1}(1-p)^{\phi_2-1} q^{\phi_3-1}(1-q)^{\phi_4-1},\end{aligned} \tag{14}$$

which can be written as

$$\mathbb{P}(s_t,p,q|z^{t-1}) \triangleq \sum_{\phi} \mathbb{P}(\phi,s_t|z^{t-1})\mathbb{P}(p,q|\phi),$$

where

$$\mathbb{P}(\phi,s_t|z^{t-1}) \triangleq \frac{C(\phi,S(z^{t-1}),s_t)\Pi_{i=1}^4\Gamma(\phi_i)}{\mathbb{P}(z^{t-1})\Gamma(\phi_1+\phi_2)\Gamma(\phi_3+\phi_4)}.$$

Therefore, the posterior $\mathbb{P}(s_t,p,q|z^{t-1})$ can be seen as a probability distribution over the energy state s_t and the posterior count ϕ. Furthermore, the posterior can be fully described by each appearance count C associated with the posterior count ϕ and the energy state s_t, up to the normalization term $\mathbb{P}(z^{t-1})$.

When we obtain a new observation z_t at time t, the posterior at time $t+1$ is updated recursively as follows

$$\begin{aligned}&\mathbb{P}(s_{t+1},p,q|z^t) = \mathbb{P}(s_{t+1},p,q|z^{t-1},z_t)\\ &= \sum_{s_t} \mathbb{P}(s_t,p,q,s_{t+1}|z^{t-1},z_t)\\ &= \sum_{s_t} \mathbb{P}(s_t,p,q,s_{t+1},z_t|z^{t-1})/\mathbb{P}(z_t|z^{t-1})\\ &= \sum_{s_t} \mathbb{P}(s_t,p,q|z^{t-1})\mathbb{P}(s_{t+1},z_t|s_t,p,q,z^{t-1})/\mathbb{P}(z_t|z^{t-1})\\ &= \sum_{s_t} \mathbb{P}(s_t,p,q|z^{t-1})\mathbb{P}(s_{t+1},z_t|s_t,p,q)/\mathbb{P}(z_t|z^{t-1}),\end{aligned} \tag{15}$$

where $\mathbb{P}(z_t|z^{t-1})$ is the normalization term.

If we harvest and observe the exact state, the total number of possible posterior counts will remain the same. For example, if we harvest and observe that $z_t = G$, this implies that $s_t = G$.

The posterior for $s_{t+1} = B$ is then

$$\begin{aligned}
&\mathbb{P}(B,p,q|z^t)\mathbb{P}(z_t|z^{t-1}) = \mathbb{P}(G,p,q|z^{t-1})\mathbb{P}(B|G,p,q)\\
&= \sum_{\phi} C(\phi, S(z^{t-1}), G)p^{\phi_1-1}(1-p)^{\phi_2-1}q^{\phi_3-1}(1-q)^{\phi_4-1}\\
&\qquad \cdot p/\mathbb{P}(z^{t-1})\\
&= \sum_{\phi} C(\phi, S(z^{t-1}), G)p^{\phi_1}(1-p)^{\phi_2-1}q^{\phi_3-1}(1-q)^{\phi_4-1}\\
&\qquad /\mathbb{P}(z^{t-1}),
\end{aligned}$$

where the second equality follows from (14). Letting $\phi' := [\phi_1+1, \phi_2, \phi_3, \phi_4]$, we have

$$\begin{aligned}
&\mathbb{P}(B,p,q|z^t)\mathbb{P}(z_t)\\
&= \sum_{\phi} C(\phi, S(z^{t-1}), G)p^{\phi_1}(1-p)^{\phi_2-1}q^{\phi_3-1}(1-q)^{\phi_4-1}\\
&= \sum_{\phi'} C(\phi, S(z^{t-1}), G)p^{\phi'_1-1}(1-p)^{\phi'_2-1}q^{\phi'_3-1}(1-q)^{\phi'_4-1}.
\end{aligned}$$

This update has the simple form that for each posterior count we increase ϕ_1 by one and keep the same appearance count.

On the other hand, the total number of possible posterior counts will be at most multiplied by two for the sleeping action, i.e., it grows exponentially with the number of sleeping actions.

For example, if the action is to sleep, i.e., $z_t = Z$, then we have to iterate over two possible states at time t since we do not know the exact state. The posterior for $s_{t+1} = B$ is then

$$\begin{aligned}
&\mathbb{P}(B,p,q|z^t)\mathbb{P}(z_t|z^{t-1}) = \sum_{s_t\in\{G,B\}} \mathbb{P}(s_t,p,q|z^{t-1})\mathbb{P}(B|s_t,p,q)\\
&= \Big[\sum_{\phi} C(\phi, S(z^{t-1}), G)p^{\phi_1-1}(1-p)^{\phi_2-1}q^{\phi_3-1}(1-q)^{\phi_4-1}\cdot p\\
&\quad + \sum_{\phi} C(\phi, S(z^{t-1}), B)p^{\phi_1-1}(1-p)^{\phi_2-1}q^{\phi_3-1}(1-q)^{\phi_4-1}\\
&\qquad \cdot (1-q)\Big]/\mathbb{P}(z^{t-1}).
\end{aligned}$$

The appearance count update can be computed in an analogous fashion, and the updates in the other scenarios can be defined similarly. An example of the update of the appearance count is shown in Figure 3. Note that two previously different posterior counts could

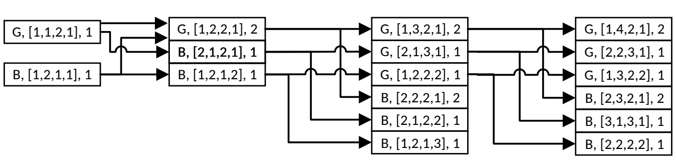

Fig. 3. A belief-update example after *two sleeping actions* and *one harvesting action with good state observation*. The numbers in the rectangle denote respectively the energy state (G or B), the posterior count ϕ and the appearance count C.

lead to the same value after one update, in which we simply add their appearance count.

B. Extended POMDP formulation of the Bayesian framework

The problem is now to derive a policy which maximizes the expected reward based on the current posterior distribution of the energy states and the state transition probabilities, obtained via the Bayesian framework described above. In [20], it has been shown that this is equivalent to deriving an optimal policy for an extended POMDP.

In what follows, we will describe how this extended POMDP is constructed. The state space of the POMDP is $\{G, B\} \times \mathbb{Z}_+^4$, i.e., the energy state and the posterior count ϕ of the Beta distribution. The action space and the reward function are the same as for the original POMDP. For brevity, we let $I_t \triangleq \{s_{t-1}, \phi, a_t\}$. By the formula of conditional probability and the independence assumptions, the joint state transition and observation probability is

$$\begin{aligned}
\mathbb{P}(s_t, \phi', z_t|I_t) &= \mathbb{P}(s_t|I_t)\mathbb{P}(z_t|I_t, s_t)\mathbb{P}(\phi'|I_t, s_t, z_t)\\
&= \mathbb{P}(s_t|s_{t-1}, \phi)\mathbb{P}(z_t|s_t)\mathbb{P}(\phi'|s_{t-1}, \phi, s_t), \quad (16)
\end{aligned}$$

where $\mathbb{P}(z_t|s_t) = 1$ if $z_t = s_t$, and $\mathbb{P}(\phi'|s_{t-1}, \phi, s_t) = 1$ if the change of state from s_{t-1} to s_t leads to the corresponding update of ϕ to ϕ'. Lastly, the transition $\mathbb{P}(s_t|s_{t-1}, \phi)$ is derived from the average p and q associated with the posterior count ϕ. For example, if $s_{t-1} = G$ and $s_t = B$, then $\mathbb{P}(s_t|s_{t-1}, \phi) = \phi_1/(\phi_1 + \phi_2)$. Therefore, the problem of deriving the optimal policy in the Bayesian framework can be solved using similar techniques as we used for the original POMDP. The optimal policy tackles the exploration and exploitation dilemma by incorporating the uncertainty in the transition probabilities in the decision making processes.

C. A heuristic learning algorithm based on posterior sampling

It is computationally difficult to solve the extended POMDP exactly due to its large state space. More precisely, during the Bayesian update, we keep the appearance count of all possible posterior counts ϕ and all possible energy states (G or B). The challenge is that the number of possible posterior counts ϕ is multiplied by two after the sleeping action, and it can grow to infinity. One approach could be to ignore the posterior update with the sleeping action, in which case the number of posterior counts is kept constant at two. However, this is equivalent to assuming that the unknown energy state does not change during the sleeping period.

Instead, we propose the heuristic posterior sampling, Algorithm 2, inspired by the discussion of approximate belief monitoring in [20, Section 5.1] and the posterior sampling algorithm proposed in [27]. The basic idea is to keep the K posterior counts that have the largest appearance count in the Bayesian update. If the energy state was in good state, then we keep harvesting. If the energy state was in bad state, then we get a sample of transition probabilities from the posterior distributions, and find the optimal sleeping time corresponding to the sampled transition probabilities. The idea leverages on the fact the optimal policy with respect to a given set of transition probabilities is threshold-based and can be pre-computed off-line.

More precisely, the algorithm maintains the value $\psi^G \triangleq [\phi_1, \phi_2, \phi_3, \phi_4, n]$ that denotes the appearance count n that leads to the posterior count $[\phi_1, \phi_2, \phi_3, \phi_4]$ and the good state. The value ψ^B is defined similarly. The two procedures in Line 24 and Line 26 show the computation of the update of the posterior count and appearance count with good and bad state observations, respectively. We uniformly pick a posterior count according to their appearance counts shown in Line 11 to reduce computational complexity. The transition probability is taken to be the mean of the Beta distribution corresponding to the sampled posterior count as shown in Line 12. Lastly, with the sleeping action, we have to invoke both good state and bad state updates in Line 18 and 19, since the state is not observed. Note that both the good state update and the bad state update take constant time, and there are at most $2K$ posterior counts in the heuristic algorithm. Hence, at each time slot, the complexity of Algorithm 2 is in the order of $\mathcal{O}(K)$.

V. Numerical Examples

A. Known transition probabilities

In the case of known transition probabilities of the Markov chain model, the optimal energy harvesting policy can be fully characterized

Algorithm 2: Posterior-sampling algorithm

Input: r, γ, K, optimal policy lookup table

Initialization: Let sleeping time $w = 0$

while *true* **do**

  if *sleeping time* $w = 0$ **then**

    Harvest energy

    if *Successfully with good state* **then**

      `Good State Update()`

      Sleeping time $w = 0$

    else

      `Bad State Update()`

      repeat

        Draw ψ^G or ψ^B randomly proportional to the count n

        Let $p = \phi_1/(\phi_1 + \phi_2), q = \phi_3/(\phi_3 + \phi_4)$

      until $1 - p > q$

      Find sleeping time w from the lookup table

    end

  else

    Sleep and decrease sleeping time $w = w - 1$

    `Good State Update()`

    `Bad State Update()`

  end

  Merge $\overline{\psi^G}$ and $\overline{\psi^B}$ with same posterior count by summing appearance count n

  Assign $2K$ items of $\overline{\psi^G}$ and $\overline{\psi^B}$ with the highest number of n to ψ^G and ψ^B, respectively.

end

Procedure `Good State Update()`

  For each ψ^G, generate new $\overline{\psi^G}$ such that $\overline{\psi^G}(\phi_2) = \psi^G(\phi_2) + 1$ and new $\overline{\psi^B}$ such that $\overline{\psi^B}(\phi_1) = \psi^G(\phi_1) + 1$

Procedure `Bad State Update()`

  For each ψ^B, generate new $\overline{\psi^G}$ such that $\overline{\psi^G}(\phi_3) = \psi^G(\phi_3) + 1$ and new $\overline{\psi^B}$ such that $\overline{\psi^B}(\phi_4) = \psi^G(\phi_4) + 1$

by the sleeping time after an unsuccessful harvesting attempt (cf. Proposition 1). For different values of reward and cost, we show in Figure 4–6 the optimal sleeping time, indexed by the average number of time-slots the model stays in the bad harvesting state $T_B = 1/q$ and the probability of being in the good state $\Pi_G = q/(p+q)$. Note that the bottom-left region without any color corresponds to the case $1-p < q$. The region with black color denotes the scenario in which it is not optimal to harvest any more after an unsuccessful harvesting.

From these figures, we first observe the natural monotonicity of longer sleeping time with respect to longer burst lengths and smaller success probabilities. Moreover, the optimal sleeping time depends not only on the burst length and the success probability, but also depends on the ratio between the reward r_1 and the penalty r_0. One might be mislead to believe that if the reward is much larger than the cost, then the optimal policy should harvest all the time. However, Figure 4 shows that for a rather large parameter space, the optimal policy is to sleep for one or two time-slots after an unsuccessful harvesting. On the other hand, when the cost is larger (i.e. larger r_0), it is better not to harvest at all in a larger parameter space. Nevertheless, there still exists a non-trivial selection of the sleeping time to maximize the total harvested energy as shown in Figure 6. Figure 7 shows that the accumulated energy can be significant.

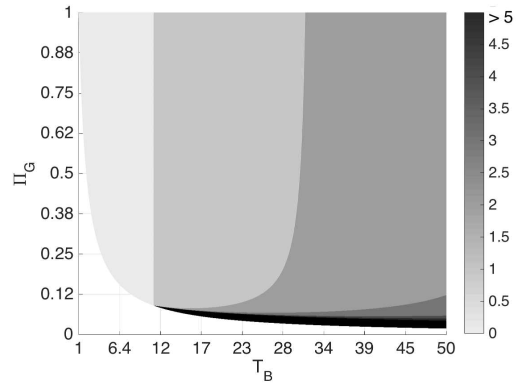

Fig. 4. Optimal sleeping time with $r_1 = 10$, $r_0 = 1$ and $\gamma = 0.99$.

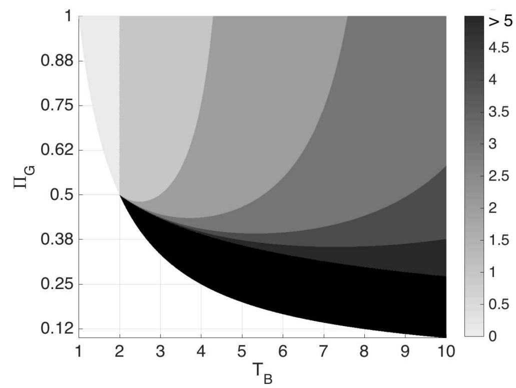

Fig. 5. Optimal sleeping time with $r_1 = 10$, $r_0 = 10$ and $\gamma = 0.99$.

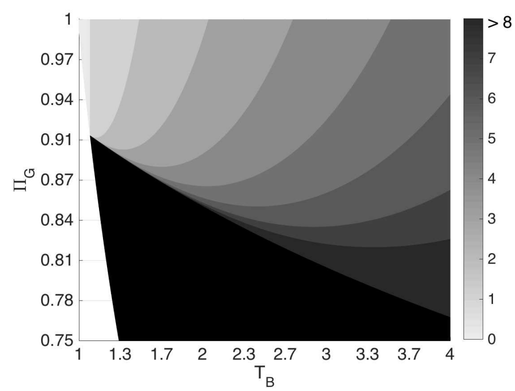

Fig. 6. Optimal sleeping time with $r_1 = 1$, $r_0 = 10$ and $\gamma = 0.99$.

In these numerical examples, we let the reward r_1 and the penalty r_0 be close, and the ratio is between 0.1 and 10. We believe that such choices are practical. For example, in AT86RF231 [28] (a leading low power radio transceiver), channel sensing consumes $3\mu J$ of energy since one clear channel assessment takes $140\mu s$ and the energy cost for keeping the radio on is $22mW$. Moreover, the energy harvesting rate of the current technology is around $200\mu W$ [1], [29]. Suppose that the coherence time of the energy source is T milliseconds, which corresponds to the duration of the time-slot. The ratio r_1/r_0 is roughly $(0.2T - 3)/3$, and it ranges from 0.3 to 10 if $T \in [20, 200]$ milliseconds. Therefore, the ratio between the reward r_1 and the penalty r_0 is neither too large nor too small, and the POMDP and the threshold-based optimal policies are very useful in practice to derive

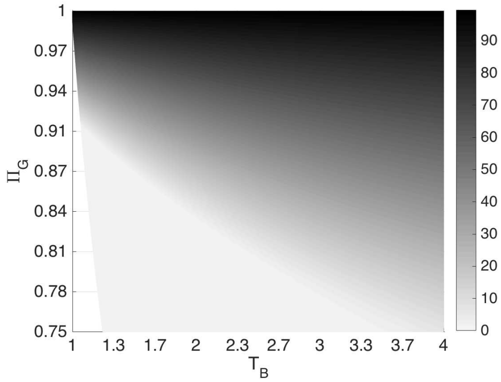

Fig. 7. Maximum harvested energy with $r_1 = 1$, $r_0 = 10$ and $\gamma = 0.99$.

the non-trivial optimal sleeping times.

Furthermore, the accumulated energy is sufficient for an ultra low-power device even in the worst case scenario shown in Figure 7. The reward r_1 could be on the order of micro joules, with the energy harvesting rate in around $200\mu W$ [1], [29] and the coherence time of the energy source in dozens of milliseconds. The total *discounted* reward is on the order of dozens of micro joules shown in Figure 7. This accumulated energy is sufficient in the scenario where devices operate most of the time in standby mode, where the energy consumption can be in the order of micro watts [28], [30].

In Figure 8, we compare the performance of the POMDP-based solution with the performance of some other policies. All these policies obtain a larger reward with a higher probability of being in the good state Π_G. The genie policy is the best among them, since it assumes that the state is known before harvesting. We can also observe that the POMDP based solution always outperforms the two heuristic policies of harvesting all the time and sleeping one time slot after an unsuccessful attempt.

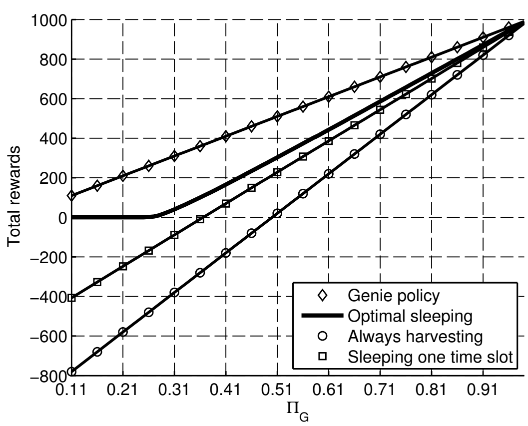

Fig. 8. Comparison with other policies with $r_1 = 10, r_0 = 10, \gamma = 0.99$ and $T_B = 10$.

Recall that the threshold-based optimal policy in Proposition 1 induces a discrete-time Markov chain with state $(S, \mathcal{E})$ denoting the energy arrival state at the previous time-slot and the energy level at the current time-slot, respectively. For battery-operated devices, once the battery is completely depleted, we cannot turn on the radio to harvest anymore, which corresponds to the absorbing states $(S, 0)$ for any S in this Markov chain. If the battery has a maximum capacity $\overline{\mathcal{E}}$, this introduces another set of absorbing states $(S, \overline{\mathcal{E}})$ for any S. Without loss of generality, we assume the energy level in the battery is a multiple of the harvested energy at each time-slot and the cost for an unsuccessful harvesting. Hence, this Markov chain has a finite number of states, and we can derive some interesting parameters by standard analysis tools from the absorbing Markov chain theory [31].

Figure 9 shows the full-charge probability under a hypothetical energy harvesting device with average success energy arrival probability equal 0.7 and under different initial energy levels. We assume that the maximum battery level is 100 units, and one successful harvesting accumulates one unit of energy while one unsuccessful harvesting costs one unit of energy. The plots can guide us in designing appropriate packet transmission policies. For example, for the case of burst length equal 10, we should restrain from transmitting the packet once the battery is around 20% full if we want to keep the depletion probability smaller than $5 \cdot 10^{-4}$.

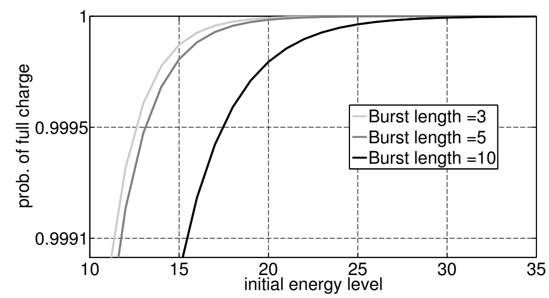

Fig. 9. The full-charge probability under different initial energy levels and average burst length.

Lastly, Figure 10 shows the average number of time-slots to reach full-charge (if the device manages to fully charge the battery) under different initial energy levels and average burst lengths. The figure shows a decreasing and almost linear relation between the initial energy level and the average number of time-slots when the initial energy level becomes larger. Similarly, the slope of these numbers can help us determine whether we can expect to be able to support a sensor application with a specified data transmission rate. Suppose the cost for one packet transmission is 40. If the data rate is larger than one packet per 50 time-slots, the energy harvesting device would quickly deplete the battery, since it takes more than 50 time-slots to harvest 40 units of energy. On the other hand, if the data rate is smaller than one packet per 100 time-slots, then we are confident that it can support such applications.

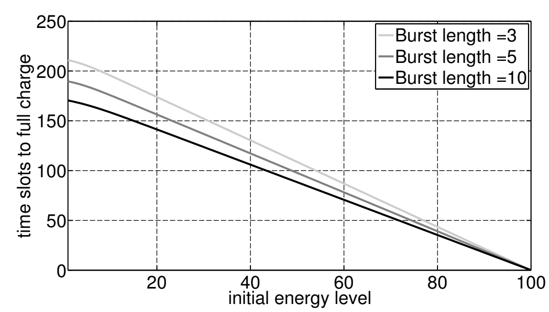

Fig. 10. The expected number of time-slots to reach full-charge under different initial energy levels and average burst length.

B. Unknown transition probabilities

In this section, we demonstrate the performance of the Bayesian learning algorithm. Figure 11 shows that the performance of Algorithm 2 outperforms other heuristic learning algorithms in terms of the total discounted reward. The results are averaged over three hundred independent energy arrival sample paths generated from the unknown Markov chain, and for each sample path the rewards are averaged over one hundred independent runs. The accumulated total reward is computed by summing over all received rewards before the given time slot. Furthermore, the received reward at each time slot is discounted from time slot 0, i.e., the reward received at time slot t is discounted by γ^t. Note that at each time slot, we evaluate the accumulated total rewards during the learning phase before that time, not the expected total rewards given the learned parameters p and q.

In the heuristic posterior sampling method, the posterior count is only updated when we have an observation of the state transition (i.e., two consecutive harvesting actions that both reveal the state of the Markov chain). In the heuristic random sampling method, we replace Line 11 and Line 12 in Algorithm 2 with a uniformly selected set of parameters p and q. Because of the heuristic choice of keeping only K posterior counts, the Bayesian update is not exact and the parameter estimation is biased. For example, we observe that $T_B = 1.9$ and $\Pi_G = 0.7$ for the heuristic posterior sampling method in Figure 11. However, its total reward still outperforms others as a result of its smarter exploration decisions during the learning phase. Note also that due to the discount factor γ being strictly smaller than one, the reward and the penalty after five hundred time-slots are negligible compared to the already accumulated rewards. In summary, the purpose of the posterior sampling method is not to estimate the parameters, but to achieve good performance with unknown parameters. For example, if the environment changes and it is assumed to be "block-wise stationary", one can re-start the learning phase after a coherent period of time with this posterior sampling method.

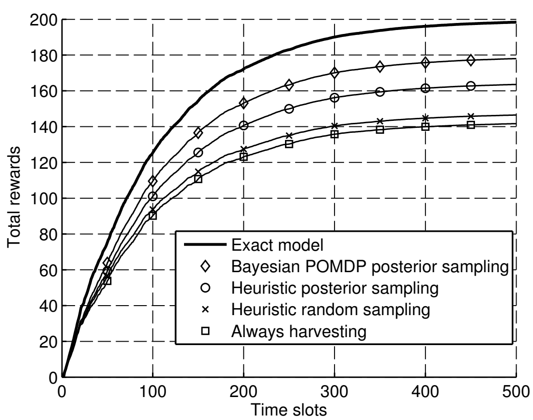

Fig. 11. Total rewards with different algorithms with $\Pi_G = 0.6$, $T_B = 2.5$, $r_0 = 10$, $r_1 = 10$, $\gamma = 0.99$, $K = 20$.

VI. Conclusions and Future Work

A. Conclusions

In this paper, we studied the problem of deciding when a wireless node with RF-EH capabilities should try and harvest ambient RF energy and when it should sleep instead. We assumed that the overall energy level is constant during one time-slot, and may change in the next time-slot according to a two-state Gilbert-Elliott Markov chain model. Based on this model, we considered two cases: first, we have knowledge of the transition probabilities of the Markov chain. On these grounds, we formulated the problem as a Partially Observable Markov Decision Process (POMDP) and determined a threshold-based optimal policy. Second, we assumed that we do not have any knowledge about these parameters and formulated the problem as a Bayesian adaptive POMDP. To simplify computations, we also proposed a heuristic posterior sampling algorithm. Numerical examples have shown the benefits of our approach.

B. Future Work

In this work, we considered a model in which RF energy is positively correlated in time. Part of ongoing work is to extend our methodologies to account for the cases for which RF energy is not positively correlated in time. Furthermore, since the process may result in different energy intakes or losses, we will extend our model to allow for random energy levels. Our analysis has also assumed that the parameters of the underlying Markov chain are fixed throughout; however, in reality these parameters might change due to the possibly changing environment and thus part of our current research focuses on investigating how our approach can be made robust to such changes.

Going one step further, to make the model even more realistic, one should consider the battery stage and capture the degradation status of the battery (see, for example, [32, § VII]) to develop policies that are aware of the degradation of the battery.

Finally, the problem of harvesting from multiple channels is of high interest when considering multi-antenna devices. The formulation of this problem falls into the restless bandit problem framework and it is left for future work.

Appendix A
Supporting lemmas

Lemma 1 is a well-known result in the POMDP literature [33]. It characterizes the piecewise linear convex segments of the value function in Equation (17) and (18), and is useful for the proofs of other lemmas and propositions.

Lemma 1. *The value function $V_t(b)$ in the value iteration algorithm at any time t is a piecewise linear convex function over belief b, i.e.,*

$$V_t(b) = \max_{\{\alpha,\beta\}\in\Gamma_t\subset\mathbb{R}^2} \{\alpha + \beta b\},$$

where the set Γ_t is computed iteratively from the set Γ_{t-1} with the initial condition $\Gamma_0 = \{0, 0\}$.

Proof: We prove the lemma by induction on time t. The statement is correct when $t = 0$ with $\Gamma_0 = \{0, 0\}$ since $V_0(b) = 0$ for all b. Suppose the statement is correct for any t. By Equation (6), the value function of sleeping action at time $t + 1$ is

$$\begin{aligned}Q_{t+1}(b, \mathcal{S}) &= \gamma V_t(q + b(1-p-q)) \\ &= \gamma \max_{\{\alpha,\beta\}\in\Gamma_t} \{\alpha + \beta(q + b(1-p-q))\} \\ &= \max_{\{\alpha,\beta\}\in\Gamma_t} \{\gamma(\alpha + \beta q) + b\gamma\beta(1-p-q)\}.\end{aligned}$$

Define

$$\begin{aligned}\Gamma_{s,t+1} &\triangleq \{\gamma(\alpha + \beta q), \gamma\beta(1-p-q) : \forall\{\alpha,\beta\} \in \Gamma_t\}, \\ \alpha_s &\triangleq \gamma(\alpha + \beta q), \\ \beta_s &\triangleq \gamma\beta(1-p-q).\end{aligned}$$

Hence, we have

$$Q_{t+1}(b, \mathcal{S}) = \max_{\{\alpha_s,\beta_s\}\in\Gamma_{s,t+1}} \{\alpha_s + \beta_s b\}. \tag{17}$$

The value function of the harvesting action is

$$\begin{aligned}Q_{t+1}(b, \mathcal{H}) \triangleq\ & (r_0 + r_1)b - r_0 + \gamma V_t(q)(1-b) + \gamma V_t(1-p)b \\ &= -r_0 + \gamma V_t(q) + (r_0 + r_1 + \gamma(V_t(1-p) - V_t(q)))b.\end{aligned}$$

Define

$$\begin{aligned}\alpha_{h,t} &\triangleq -r_0 + \gamma V_t(q), \\ \beta_{h,t} &\triangleq r_0 + r_1 + \gamma(V_t(1-p) - V_t(q)).\end{aligned}$$

We then have

$$Q_{t+1}(b, \mathcal{H}) = \alpha_{h,t} + \beta_{h,t} b. \tag{18}$$

Since $V_{t+1}(b) = \max\{Q_{t+1}(b, \mathcal{S}), Q_{t+1}(b, \mathcal{H})\}$, the statement is proved by defining $\Gamma_{t+1} \triangleq \{\alpha_{h,t}, \beta_{h,t}\} \bigcup \Gamma_{s,t+1}$. ■

Lemma 2. *For any t, if $b_1 \geq b_2$, then $V_t(b_1) \geq V_t(b_2)$. For any $\{\alpha, \beta\} \in \Gamma_t$, we have $\beta \geq 0$.*

Proof: We prove the proposition by induction on time t. Since $V_0(b) = 0$ for all b at time $t = 0$ and $\Gamma_0 = \{0, 0\}$, the statement is correct at time $t = 0$. Suppose the statement is correct at time t. Since $1 - p - q \geq 0$ and $\beta \geq 0$, we have that

$$\gamma(\alpha + \beta q) + b_1\gamma\beta(1 - p - q) \geq \gamma(\alpha + \beta q) + b_2\gamma\beta(1 - p - q). \tag{19}$$

This means that for any given pair of α_s and β_s, we have that $\alpha_s + \beta_s b_1 \geq \alpha_s + \beta_s b_2$. By Equation (17), we have $Q_{t+1}(b_1, \mathcal{S}) \geq Q_{t+1}(b_2, \mathcal{S})$. Since $1 - p > q$, we also have $V_t(1 - p) \geq V_t(q)$ by the induction condition, and hence $\beta_{h,t} \geq 0$. By Equation (18), we have $Q_{t+1}(b_1, \mathcal{H}) \geq Q_{t+1}(b_2, \mathcal{H})$. Hence, we have that $V_{t+1}(b_1) \geq V_{t+1}(b_2)$. Similarly, we can also derive that $\beta \geq 0$ for any $\{\alpha, \beta\} \in \Gamma_{t+1}$. ■

APPENDIX B
PROOF OF PROPOSITION 1

The proof relies on two Lemmas presented in Appendix A. We first prove that the optimal action is to sleep for any belief $b < \bar{b}$ and to harvest for any belief $b \geq \bar{b}$. The definition of $\bar{b}$ implies that for any $b < \bar{b}$, we have that $Q_\infty(b, \mathcal{H}) < Q_\infty(b, \mathcal{S})$, and hence it is always optimal to sleep for belief $b < \bar{b}$.

If we let $t \to \infty$ in Equation (17) and Equation (18), then we have that

$$Q_\infty(b, \mathcal{H}) = \alpha_{h,\infty} + \beta_{h,\infty} b,$$
$$Q_\infty(b, \mathcal{S}) = \max_{\{\alpha_s, \beta_s\} \in \Gamma_{s,\infty}} \{\alpha_s + \beta_s b\},$$

where $\Gamma_{s,\infty} = \{\gamma(\alpha + \beta q), \gamma\beta(1 - p - q) : \forall\{\alpha, \beta\} \in \Gamma_\infty\}$ and $\Gamma_\infty = \Gamma_{s,\infty} \bigcup \{\alpha_{h,\infty}, \beta_{h,\infty}\}$. Let $\beta_{s'} \triangleq \max_{\{\alpha_s, \beta_s\} \in \Gamma_{s,\infty}} \{\beta_s\}$. From the definition of $\Gamma_{s,\infty}$, there exists an $\beta_{s''} \in \Gamma_\infty$ such that $\gamma(1 - p - q)\beta_{s''} = \beta_{s'}$. Suppose $\beta_{s''} \in \Gamma_{s,\infty}$. We then have that $\beta_{s''} \leq \beta_{s'} = \gamma(1 - p - q)\beta_{s''}$. Since $\beta \geq 0$ for any $\{\alpha, \beta\} \in \Gamma_t$ from Lemma 2 and $1 - p - q > 0$ and $0 < \gamma < 1$, we have a contradiction here, and thus $\beta_{s''} \notin \Gamma_{s,\infty}$. Since $\beta_{s''} \in \Gamma_\infty$, we have $\beta_{s''} = \beta_{h,\infty}$. Hence, we have

$$\beta_{h,\infty} = \beta_{s''} = \beta_{s'}/(\gamma(1 - p - q)) > \beta_{s'} \geq \max_{\{\alpha_s, \beta_s\} \in \Gamma_{s,\infty}} \{\beta_s\}.$$

Since $V_\infty(\bar{b}, \mathcal{H}) \geq V_\infty(\bar{b}, \mathcal{S})$, it follows that $V_\infty(b, \mathcal{H}) \geq V_\infty(b, \mathcal{S})$ for any $b \geq \bar{b}$. Hence, the optimal action is to harvest for any belief $b \geq \bar{b}$.

Suppose $\bar{b} < q/(p + q)$. Since the belief after a successful harvesting is $1 - p > q/(p + q)$, we continue to harvest. Observe that after an unsuccessful harvesting and sleeping additionally for $t - 1$ time-slots, the belief b is

$$q \sum_{i=0}^{t-1} (1 - p - q)^i = q\frac{1 - (1 - p - q)^t}{p + q}. \tag{20}$$

Since $1 - p - q \in (0, 1)$, this is monotonically increasing with t and converges to $q/(p + q)$. The proposition follows by deriving t such that the belief is larger than the threshold $\bar{b}$.

Suppose $q/(p + q) \leq \bar{b} \leq 1 - p$. Since the belief after a successful harvesting is $1 - p$, we continue to harvest. Similarly, observe from Equation (20) that the belief b is always smaller than $q/(p + q)$ after an unsuccessful harvesting attempt. Thus, we stop harvesting after an unsuccessful harvesting attempt.

Lastly, it is straightforward to observe that we do not harvest at all if $1 - p < \bar{b}$. □

APPENDIX C
PROOF OF PROPOSITION 2

Let π^n denote the policy that sleeps n time-slots after bad state observation, and always harvests after good state observation. By Proposition 1, the optimal policy is a type of π^n policy since $\bar{b} < q/(p + q)$. In what follows, we focus on the type of π^n policy and derive the optimal sleeping time that obtains the maximum reward.

Recall that the belief after good state observation is $1 - p$, and after bad state observation is q. The belief after bad state observation and sleeping n time-slots is

$$\tilde{b} \triangleq q \sum_{i=0}^{n} (1 - p - q)^i = q\frac{1 - (1 - p - q)^{n+1}}{p + q}.$$

At belief q, the π^n policy is to sleep for n time-slots, and thus

$$V^{\pi^n}(q) = \gamma^n V^{\pi^n}(\tilde{b}). \tag{21}$$

At belief $1 - p$, the π^n policy is to harvest, and thus

$$\begin{aligned} V^{\pi^n}(1 - p) &= (1 - p)(r_0 + r_1) - r_0 \\ &\quad + \gamma p V^{\pi^n}(q) + \gamma(1 - p)V^{\pi^n}(1 - p). \end{aligned} \tag{22}$$

At belief $\tilde{b}$, the π^n policy is also to harvest, and thus

$$\begin{aligned} V^{\pi^n}(\tilde{b}) &= \tilde{b}(r_0 + r_1) - r_0 \\ &\quad + \gamma(1 - \tilde{b})V^{\pi^n}(q) + \gamma\tilde{b}V^{\pi^n}(1 - p). \end{aligned} \tag{23}$$

By solving the above Equations (21)-(22)-(23), $V^{\pi^n}(1 - p)$ corresponds to $F(n)/G(n)$. Hence, N is the optimal sleeping time that gives the maximum reward within the set of policies defined by π^n. Since the optimal policy has this structure, the proposition is proved. □

REFERENCES

[1] L. Xiao, P. Wang, D. Niyato, D. Kim, and Z. Han, "Wireless Networks with RF Energy Harvesting: A Contemporary Survey," *IEEE Communications Surveys & Tutorials*, 2015.

[2] S. Ulukus, A. Yener, E. Erkip, O. Simeone, M. Zorzi, P. Grover, and K. Huang, "Energy harvesting wireless communications: A review of recent advances," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 3, pp. 360–381, March 2015.

[3] I. Ahmed, M. M. Butt, C. Psomas, A. Mohamed, I. Krikidis, and M. Guizani, "Survey on energy harvesting wireless communications: Challenges and opportunities for radio resource allocation," *Computer Networks*, July 2015.

[4] U. Olgun, C.-C. Chen, and J. Volakis, "Investigation of Rectenna Array Configurations for Enhanced RF Power Harvesting," *IEEE Antennas and Wireless Propagation Letters*, vol. 10, pp. 262–265, 2011.

[5] V. Liu, A. Parks, V. Talla, S. Gollakota, D. Wetherall, and J. R. Smith, "Ambient backscatter: wireless communication out of thin air," in *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 4, 2013, pp. 39–50.

[6] T. Wu and H. Yang, "On the performance of overlaid wireless sensor transmission with RF energy harvesting," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 8, pp. 1693–1705, Aug 2015.

[7] D. Gunduz, K. Stamatiou, N. Michelusi, and M. Zorzi, "Designing intelligent energy harvesting communication systems," *IEEE Communications Magazine*, vol. 52, no. 1, pp. 210–216, 2014.

[8] V. Sharma, U. Mukherji, V. Joseph, and S. Gupta, "Optimal energy management policies for energy harvesting sensor nodes," *IEEE Transactions on Wireless Communications*, vol. 9, no. 4, pp. 1326–1336, 2010.

[9] N. Michelusi, K. Stamatiou, and M. Zorzi, "Transmission policies for energy harvesting sensors with time-correlated energy supply," *IEEE Transactions on Communications*, vol. 61, no. 7, pp. 2988–3001, 2013.

[10] J. Lei, R. Yates, and L. Greenstein, "A generic model for optimizing single-hop transmission policy of replenishable sensors," *IEEE Transactions on Wireless Communications*, vol. 8, no. 2, pp. 547–551, 2009.

[11] O. Ozel, K. Tutuncuoglu, J. Yang, S. Ulukus, and A. Yener, "Transmission with energy harvesting nodes in fading wireless channels: Optimal policies," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 8, pp. 1732–1743, 2011.

[12] P. Blasco, D. Gunduz, and M. Dohler, "A learning theoretic approach to energy harvesting communication system optimization," *IEEE Transactions on Wireless Communications*, vol. 12, no. 4, pp. 1872–1882, 2013.

[13] J. Fernandez-Bes, J. Cid-Sueiro, and A. Marques, "An MDP model for censoring in harvesting sensors: Optimal and approximated solutions," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 8, pp. 1717–1729, Aug 2015.

[14] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[15] N. Michelusi, L. Badia, and M. Zorzi, "Optimal transmission policies for energy harvesting devices with limited state-of-charge knowledge," *IEEE Transactions on Communications*, vol. 62, no. 11, pp. 3969–3982, 2014.

[16] N. Jaggi, K. Kar, and A. Krishnamurthy, "Rechargeable sensor activation under temporally correlated events," *Wireless Networks*, vol. 15, no. 5, pp. 619–635, 2009.

[17] 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*, vol. 7, no. 5, pp. 895–906, 2013.

[18] L. A. Johnston and V. Krishnamurthy, "Opportunistic file transfer over a fading channel: A POMDP search theory formulation with optimal threshold policies," *IEEE Transactions on Wireless Communications*, vol. 5, no. 2, pp. 394–405, 2006.

[19] Y. Chen, Q. Zhao, and A. Swami, "Distributed spectrum sensing and access in cognitive radio networks with energy constraint," *IEEE Transactions on Signal Processing*, vol. 57, no. 2, pp. 783–797, 2009.

[20] S. Ross, J. Pineau, B. Chaib-draa, and P. Kreitmann, "A Bayesian approach for learning and planning in partially observable Markov decision processes," *The Journal of Machine Learning Research*, vol. 12, pp. 1729–1770, 2011.

[21] E. N. Gilbert, "Capacity of a burst-noise channel," *Bell system technical journal*, vol. 39, no. 5, pp. 1253–1265, 1960.

[22] E. O. Elliott, "Estimates of error rates for codes on burst-noise channels," *Bell system technical journal*, vol. 42, no. 5, pp. 1977–1997, 1963.

[23] S. Kim, R. Vyas, J. Bito, K. Niotaki, A. Collado, A. Georgiadis, and M. M. Tentzeris, "Ambient rf energy-harvesting technologies for self-sustainable standalone wireless sensor platforms," *Proceedings of the IEEE*, vol. 102, no. 11, pp. 1649–1666, 2014.

[24] H. J. Visser, H. W. Pflug, and S. Keyrouz, "Rectenna demonstrators at holst centre/imec and eindhoven university of technology," in *Antennas and Propagation (EuCAP), 2015 9th European Conference on*. IEEE, 2015, pp. 1–4.

[25] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience, 2005.

[26] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial intelligence*, vol. 101, no. 1, pp. 99–134, 1998.

[27] M. Strens, "A Bayesian framework for reinforcement learning," in *ICML*, 2000, pp. 943–950.

[28] A. AT86RF231, "Low Power 2.4 GHz Transceiver for ZigBee."

[29] Z. Popovic, E. A. Falkenstein, D. Costinett, and R. Zane, "Low-power far-field wireless powering for wireless sensors," *Proceedings of the IEEE*, vol. 101, no. 6, pp. 1397–1409, 2013.

[30] T. Instruments, "MSP430FR572x Mixed-Signal Microcontrollers."

[31] J. G. Kemeny and J. L. Snell, *Finite markov chains*. van Nostrand Princeton, NJ, 1960, vol. 356.

[32] N. Michelusi, L. Badia, R. Carli, L. Corradini, and M. Zorzi, "Energy management policies for harvesting-based wireless sensor devices with battery degradation," *IEEE Transactions on Communications*, vol. 61, no. 12, pp. 4934–4947, 2013.

[33] W. S. Lovejoy, "A survey of algorithmic methods for partially observed markov decision processes," *Annals of Operations Research*, vol. 28, no. 1, pp. 47–65, 1991.

Zhenhua Zou received the double Master degree in communication engineering from Southeast University, China and Politecnico di Torino, Italy with *summa cum laude* in 2009. He received his Ph.D. degree in telecommunications at the School of Electrical Engineering, KTH The Royal Institute of Technology, Sweden in 2014. He worked as a post-doctoral researcher at the Department of Signals and Systems at Chalmers University of Technology, Sweden in 2015. His research interest includes algorithm development for real-time communication in wireless lossy networks.

Anders Gidmark was born in Sweden in 1991. He received the M.Sc.Eng degree in engineering physics at KTH Royal Institute of Technology, Stockholm, in 2015. His research interests include green communications and control theory. He is currently a Systems Engineer at the department of automation at Eurocon Engineering AB, Örnsköldsvik, Sweden.

Themistoklis Charalambous received his B.A. and M.Eng. in Electrical and Information Sciences from Cambridge University in 2005. He pursued his Ph.D. in the Control Laboratory, of the Engineering Department, Cambridge University in 2010. He worked as a research associate at Imperial College London, as a Visiting Lecturer at the Department of Electrical and Computer Engineering, University of Cyprus, and as a post-doctoral researcher at the Automatic Control Lab of the School of Electrical Engineering at the Royal Institute of Technology (KTH). He is currently a post-doctoral researcher at the Department of Signals and Systems at Chalmers University of Technology. His research involves distributed coordination and control, distributed decision making, and control to various resource allocation problems in complex and networked systems.

Mikael Johansson received the M.Sc and Ph.D. degrees in electrical engineering from Lund University, Lund, Sweden, in 1994 and 1999, respectively.

He held postdoctoral positions at Stanford University, Stanford, CA, USA, and University of California, Berkeley, CA, USA, before joining KTH Royal Institute of Technology, Stockholm, Sweden in 2002, where he now serves as Full Professor. He has published two books and more than a hundred papers, several which are highly cited and have received recognition in terms of best paper awards. He has served on the editorial boards of *Automatica* and the IEEE TRANSACTIONS ON CONTROL OF NETWORK SYSTEMS, as well as on the program committee for several top-conferences organized by IEEE and ACM. He has played a leading role in several national and international research projects in control and communications.