Online Learning Schemes for Power Allocation in Energy Harvesting Communications

Pranav Sakulkar and Bhaskar Krishnamachari

Abstract—**We consider the problem of power allocation over one or more time-varying channels with unknown distributions in energy harvesting communications. In the single-channel case, the transmitter chooses the transmit power based on the amount of stored energy in its battery with the goal of maximizing the average rate over time. We model this problem as a Markov decision process (MDP) with transmitter as the agent, battery status as the state, transmit power as the action and rate as the reward. The average reward maximization problem can be modelled by a linear program (LP) that uses the transition probabilities for the state-action pairs and their reward values to select a power allocation policy. This problem is challenging because the uncertainty in channels implies that the mean rewards associated with the state-action pairs are unknown. We therefore propose two online learning algorithms: Linear Program of Sample Means (LPSM) and Epoch-LPSM that learn these rewards and adapt their policies over time. For both algorithms, we prove that their regret is upper-bounded by a constant. To our knowledge this is the first result showing constant regret learning algorithms for MDPs with unknown mean rewards. We also prove an even stronger result about LPSM: that its policy matches the optimal policy exactly in finite expected time. Epoch-LPSM incurs a higher regret compared to LPSM, while reducing the computational requirements substantially. We further consider a multi-channel scenario where the agent also chooses a channel in each slot, and present our multi-channel LPSM (MC-LPSM) algorithm that explores different channels and uses that information to solve the LP during exploitation. MC-LPSM incurs a regret that scales logarithmically in time and linearly in the number of channels. Through a matching lower bound on the regret of any algorithm, we also prove the asymptotic order optimality of MC-LPSM.**

Index Terms—**Contextual bandits, multi-armed bandits (MABs), online learning, energy harvesting communications, Markov decision process (MDP).**

I. INTRODUCTION

Communication systems where the transmissions are powered by the energy harvested from nature have rapidly emerged as viable options for the next-generation wireless networks [1]. These advances are promising as they help us reduce the dependence on the conventional sources of energy and the carbon footprint of our systems. The ability to harvest energy promises self-sustainability and prolonged network lifetimes that are only limited by the communication hardware rather than the energy storage. Energy can be harvested from various natural sources, such as solar, thermal, chemical, biological etc. Their technologies differ in terms of their efficiency and

harvesting capabilities, depending on the mechanisms, devices, circuits used. Since these energy sources are beyond human control, energy harvesting brings up a novel aspect of irregular and random energy source for communication. This demands a fresh look at the transmission schemes used by wireless networks. The next generation wireless networks need to be designed while keeping the irregularities and randomness in mind.

The performance of the energy harvesting communication systems is dependent on the efficient utilization of energy that is currently stored in the battery, as well as that is to be harvested over time. These systems must make decisions while keeping their impact on the future operations in mind. Such problems of optimal utilization of the available resources can be classified into offline optimization [2], [3], [4] and online optimization [5], [6], [7], [8], [4] problems. In the offline optimization problems, the transmitter deterministically knows the exact amounts of the harvested energies and the data along with their exact arrival times. These assumptions are too optimistic, since the energy harvesting communication systems are usually non-deterministic. In the online optimization problems, however, the transmitter is assumed to know the distributions or some statistics of the harvesting and the data arrival processes. It may get to know their instantaneous realizations before making decisions in each slot. The problem considered in this paper falls in the category of the online optimization problems.

In our paper, the channel gains are assumed to be i.i.d. over time with an unknown distribution and the harvested energy is assumed to be stochastically varying with a known distribution. This is based on the fact that the weather conditions are more predictable than the radio frequency (RF) channels which are sensitive to time-varying multi-path fading. In the single channel case, the transmitter has to decide its transmit power level based on the current battery status with the goal maximizing the average expected transmission rate obtained over time. We model the system as a Markov decision process (MDP) with the battery status as the state, the transmit power as the action, the rate as the reward. The power allocation problem, therefore, reduces to the average reward maximization problem for an MDP. Since the channel gain distribution is unknown, their expected rates for different power levels are also unknown. The transmitter or the agent, therefore, cannot determine the optimal mapping from the battery state to the transmit power precisely. It needs to learn these rate values over time and make decisions along the way. We cast this problem as an online learning problem over an MDP.

In the multi-channel case, the agent needs to also select

P. Sakulkar and B. Krishnamachri are with the Department of Electrical Engineering, University of Southern California, Los Angeles, CA 90089, USA (e-mail: sakulkar@usc.edu, bkrishna@usc.edu).

a channel from the set of channels for transmission. The gain-distributions of these channels are, in general, different and unknown to the agent. Due to different distributions, the optimal channels for different transmit powers can be different. The agent, therefore, needs to explore different channels over time to be able to estimate their rate values and use these estimates to choose a power-level and a channel at each time. This exploration of different channels leads the agent into making non-optimal power and channel decisions and thus hampers the performance of the online learning algorithm.

One interesting feature of this problem is that the data-rate obtained during transmission is a known function of the chosen transmit power level and the instantaneous channel condition. Whenever a certain power is used for transmission, the agent can figure out the instantaneous channel condition once the instantaneous rate is revealed to it. This information about the instantaneous gain of the chosen channel can, therefore, be used to update the rate-estimates for all power levels. The knowledge of the rate function can be used to speed up the learning process in this manner.

A. Contributions

The problem of maximizing the average expected reward of an MDP can be formulated as a linear program (LP). The solution of this LP gives the stationary distribution over the state-action pairs under the optimal policy. If the MDP is ergodic, then there exists a deterministic optimal policy. We model the problem of communication over a single channel using the harvested energy as an MDP and prove the ergodicity of the MDP under certain assumptions about the transmit power and the distribution of harvested energy. This helps us focus only on the deterministic state-action mappings which are finite in number.

The LP formulation helps us characterize the optimal policy that depends on the transition probabilities for the state-action pairs and their corresponding mean rewards. We use the optimal mean reward obtained from the LP as a benchmark to compare the performance of our algorithms with. Since the mean rewards associated the state-action pairs are unknown to the agent, we propose two online learning algorithms: LPSM and Epoch-LPSM that learn these rewards and adapt their policies along the way. The LPSM algorithm solves the LP at each step to decide its current policy based on its current sample mean estimates for the rewards, while the Epoch-LPSM algorithm divides the time into epochs, solves the LP only at the beginning of the epochs and follows the obtained policy throughout that epoch. We measure the performance of our online algorithms in terms of their regrets, defined as the cumulative difference between the optimal mean reward and the instantaneous reward of the algorithm. We prove that the reward loss or regret incurred by each of these algorithms is upper bounded by a constant. To our knowledge this is the first result where constant regret algorithms are proposed for the average reward maximization problem over MDPs with stochastic rewards with unknown means. We further prove that the LPSM algorithm starts following the genie's optimal policy in finite expected time. The finite expected time is an even stronger result than the constant regret guarantee.

Our proposed single channel algorithms greatly differ in their computational requirements. Epoch-LPSM incurs a higher regret compared to LPSM, but reduces the computational requirements substantially. LPSM solves a total of T LPs in time T, whereas Epoch-LPSM solves only $O(\ln T)$ number of LPs. We introduce two parameters n_0 and η that reveal the computation vs regret tradeoff for Epoch-LPSM. Tuning these parameters allows the agent to control the system based on its performance requirements.

We extend our framework to the case of multiple channels where the agent also needs to select the transmission channel in each slot. We present our MC-LPSM algorithm that deterministically separates exploration from exploitation. MC-LPSM explores different channels to learn their expected rewards and uses that information to solve the average reward maximization LP during the exploitation slots. The length of the exploration sequence scales logarithmically over time and contributes to the bulk of the regret. This exploration, however, helps us bound the exploitation regret by constant. We, therefore, prove a regret bound for MC-LPSM that scales logarithmically in time and linearly in the number of channels. This design of the exploration sequence, however, needs to know a lower bound on the difference in rates for the channels. We observe that this need of knowing some extra information about the system can be eliminated by using a longer exploration sequence as proposed in [9]. The regret of this design can be made arbitrarily close to the logarithmic order. We also prove an asymptotic regret lower bound of $\Omega(\ln T)$ for any algorithm under certain conditions. This proves the asymptotic order optimality of the proposed MC-LPSM approach. We further show that, similar to Epoch-LPSM, the MC-LPSM algorithm also solves only $O(\ln T)$ number of LPs in time T.

We show that the proposed online learning algorithms also work for cost minimization problems in packet scheduling with power-delay tradeoff with minor changes.

B. Organization

This paper is organized as follows. First, we describe the model for the energy harvesting communication system using a single channel, formulate this problem as an MDP and discuss the structure of the optimal policy in section III. We then propose our online learning algorithms LPSM and Epoch-LPSM for single channel systems and prove their regret bounds in section IV. We extend our approach to multi-channel systems, propose our MC-LPSM algorithm, and analyze its regret in section V. In section VI, we show that our online learning framework can also model the average cost minimization problems over MDPs. Section VII presents the results of numerical simulations for this problem and section VIII concludes the paper. We also include appendices A and B to discuss and prove some of the technical lemmas at the end of the paper.

II. RELATED WORK

The offline optimization problems in the energy harvesting communications assume a deterministic system and the exact

knowledge of energy and data arrival times and their amounts. In [2], the goal is to minimize the time by which all the packets are delivered. In [3], a finite horizon setting is considered with the goal of maximizing the amount of transmitted data. These two problems are proved to be duals of each other in [3]. In the online problems, the system is usually modelled as an MDP with the objective being the maximization of the average reward over time. These problems make specific distributional assumptions about the energy harvesting process and the channel gains. In the Markovian setting considered in [5], the energy replenishment process and the packet arrival process are assumed to follow Poisson distributions. Each packet has a random value assigned to it and the reward, in this setting, corresponds to the sum of the values of the successfully transmitted packets. In [6], the power management problem is formulated as an MDP, where the transmitter is assumed to know the full channel state information before the transmission in each slot. The properties of the optimal transmission policy are characterized using dynamic programming for this setting. In [7], the energy harvesting process and the data arrivals are assumed to follow Bernoulli distributions and a simple AWGN channel is assumed for transmission. A policy iteration based scheme is designed to minimize the transmission errors for the system. In [8], power allocation policies over a finite time horizon with known channel gain and harvested energy distributions are studied. In [10], [11], power control policies are analyzed for communication systems with retransmissions. These MDP settings assume that energy harvesting follows a stationary Bernoulli process. In [10], the channel is assumed to follow a known fading model and in [11], the packet error rate is assumed to be known. In [4], the offline and online versions of the throughput optimization problem are studied for a fading channel and Poisson energy arrival process. In our paper, however, we do not assume specific distributions for the energy arrival process and the channel gain-to-noise ratios.

Our problem can be seen from the lens of contextual bandits, which are extensions of the standard multi-armed bandits (MABs). In the standard MAB problem [12], [13], [9], the agent is presented with a set of arms each providing stochastic rewards over time with unknown distributions and it has to choose an arm in each trial with the goal of maximizing the sum reward over time. In [12], Lai and Robbins provide an asymptotic lower bound of $\Omega(\ln T)$ on the expected regret of any algorithm for this problem. In [13], an upper confidence bound based policy called UCB1 and a randomized policy that separates exploration from exploitation called ϵ-greedy are proposed and are also proved to achieve logarithmic regret bounds for arm distributions with finite support. In [9], a deterministic equivalent of ϵ-greedy called DSEE is proposed and proved to provide similar regret guarantees. In the contextual bandits, the agent also sees some side-information before making its decision in each slot. In the standard contextual bandit problems [14], [15], [16], [17], [18], the contexts are assumed to be drawn from an unknown distribution independently over time. In our problem, the battery state can be viewed as the context. We model the context transitions by an MDP, since the agent's action at

time t affects not only the instantaneous reward but also the context in slot $t + 1$. The agent, therefore, needs to decide the actions with the global objective in mind, i.e. maximizing the average reward over time. The algorithms presented in [15], [16], [17] do not assume any specific relation between the context and the reward. The DCB(ϵ) algorithm presented in [18], however, assumes that the mapping from the context and random instance to the reward is a known function and uses this function knowledge to reduce the expected regret. It must, however, be noted that the MDP formulation generalizes the contextual bandit setting in [18], since the i.i.d. context case can be viewed as a single state MDP.

Our problem is also closely related to the reinforcement learning problem over MDPs from [19], [20], [21]. The objective for these problems is to maximize the average undiscounted reward over time. In [19], [20], the agent is unaware of the transition probabilities and the mean rewards corresponding to the state-action pairs. In [21], the agent knows the mean rewards, but the transition probabilities are still unknown. In our problem, the mean rewards are unknown, while the transition probabilities of the MDP can be inferred from the knowledge of the arrival distribution and the action taken from each state. In contrast to the works above, for our problem motivated by the practical application in energy harvesting communications, we show that the learning incurs a constant regret in the single channel case.

III. System Model

We describe the model of the energy harvesting communication system considered in this paper using a single channel. Consider a time-slotted energy harvesting communication system where the transmitter uses the harvested power for transmission over a channel with stochastically varying channel gains with unknown distribution as shown in figure 1. We further assume that the channel gain-to-noise ratio is i.i.d. over time. Let p_t denote the harvested power in the t-th slot which is assumed to be i.i.d. over time. Let Q_t denote the stored energy in the transmitter's battery that has a capacity of $Q_{\max}$. Assume that the transmitter decides to use $q_t (\leq Q_t)$ amount of power for transmission in t-th slot. We assume discrete and finite number of power levels for the harvested and transmit powers. The rate obtained during the t-th slot is assumed to follow a relationship

$$r_t = B \log_2(1 + q_t X_t), \tag{1}$$

where X_t denotes the instantaneous channel gain-to-noise ratio of the channel which is assumed to be i.i.d. over time and B is the channel bandwidth. The battery state gets updated in the next slot as

$$Q_{t+1} = \min\{Q_t - q_t + p_t, Q_{\max}\}. \tag{2}$$

The goal is to utilize the harvested power and choose a transmit power q_t in each slot sequentially to maximize the expected average rate $\lim_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=1}^{T} r_t\right]$ obtained over time.
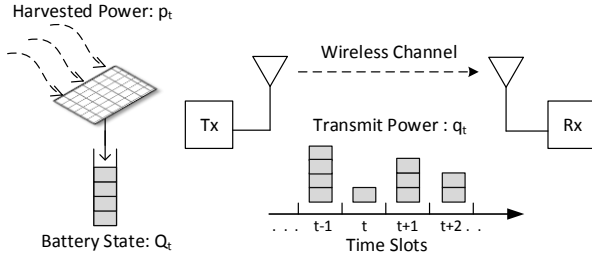
Fig. 1. Power allocation over a wireless channel in energy harvesting communications

A. Problem Formulation

Consider an MDP $\mathcal{M}$ with a finite state space $\mathcal{S}$ and a finite action space $\mathcal{A}$. Let $\mathcal{A}_s \subset \mathcal{A}$ denote the set of allowed actions from state s. When the agent chooses an action $a_t \in \mathcal{A}_{s_t}$ in state $s_t \in \mathcal{S}$, it receives a random reward $r_t(s_t, a_t)$. Based on the agent's decision the system undergoes a random transition to a state s_{t+1} according to the transition probability $P(s_{t+1} \mid s_t, a_t)$. In the energy harvesting problem, the battery status Q_t represents the system state s_t and the transmit power q_t represents the action taken a_t at any slot t.

In this paper, we consider systems where the random rewards of various state action pairs can be modelled as

$$r_t(s_t, a_t) = f(s_t, a_t, X_t), \qquad (3)$$

where f is a reward function known to the agent and X_t is a random variable internal to the system that is i.i.d. over time. Note that in the energy harvesting communications problem, the reward is the rate obtained at each slot and the reward function is defined in equation (1). In this problem, the channel gain-to-noise ratio X_t corresponds to the system's internal random variable. We assume that the distribution of the harvested energy p_t is known to the agent. This implies that the state transition probabilities $P(s_{t+1} \mid s_t, a_t)$ are inferred by the agent based on the update equation (2).

A policy is defined as any rule for choosing the actions in successive time slots. The action chosen at time t may, therefore, depend on the history of previous states, actions and rewards. It may even be randomized such that the action $a \in \mathcal{A}_s$ is chosen from some distribution over the actions. A policy is said to be stationary, if the action chosen at time t is only a function of the system state at t. This means that a deterministic stationary policy β is a mapping from the state $s \in \mathcal{S}$ to its corresponding action $a \in \mathcal{A}_s$. When a stationary policy is played, the sequence of states $\{s_t \mid t = 1, 2, \cdots\}$ follows a Markov chain. An MDP is said to be *ergodic*, if every deterministic stationary policy leads to an irreducible and aperiodic Markov chain. According to section V.3 from [22], the average reward can be maximized by an appropriate deterministic stationary policy β^* for an ergodic MDP with finite state space. In order to arrive at an ergodic MDP for the energy harvesting communications problem, we make following assumptions:

AS-1 When $Q_t > 0$, the transmit power $q_t > 0$.

AS-2 The distribution of the harvested energy is such that $\Pr\{p_t = p\} > 0$ for all $0 \leq p \leq Q_{\max}$.

Proposition 1. *Under assumptions AS-1 and AS-2, the MDP corresponding to the transmit power selection problem in energy harvesting communications is ergodic.*

Proof: Consider any policy β and let $P^{(n)}(s, s')$ be the n-step transition probabilities associated with the Markov chain resulting from the policy.

First, we prove that $P^{(1)}(s, s') > 0$ for any $s' \geq s$ as follows. According to the state update equations,

$$s_{t+1} = s_t - \beta(s_t) + p_t. \qquad (4)$$

The transition probabilities can, therefore, be expressed as

$$P^{(1)}(s, s') = \Pr\{p = s' - s + \beta(s)\} \geq 0, \qquad (5)$$

since $s' \geq s$ and $\beta(s) \geq 0$ for all states. This implies that any state $s' \in \mathcal{S}$ is accessible from any other state s in the resultant Markov chain, if $s \leq s'$.

Now, we prove that $P^{(1)}(s, s-1) > 0$ for all $s \geq 1$ as follows. From equation (5), we observe that

$$P^{(1)}(s, s-1) = \Pr\{p = \beta(s) - 1\} \geq 0, \qquad (6)$$

since $\beta(s) \geq 1$ for all $s \geq 1$. This implies that every state $s \in \mathcal{S}$ is accessible from the state $s+1$ in the resultant Markov chain.

Equations (5) and (6) imply that all the state pairs $(s, s+1)$ communicate with each other. Since communication is an equivalence relationship, all the states communicate with each other and the resultant Markov chain is irreducible. Also, equation (5) implies that $P^{(1)}(s, s) > 0$ for all the states and the Markov chain is, therefore, aperiodic. ∎

Under assumptions AS-1 and AS-2, the MDP under consideration is ergodic and we can restrict ourselves to the set of deterministic stationary policies which we interchangeably refer to as policies henceforth. Let $\mu(s, a)$ denote the expected reward associated with the state-actions pair (s, a) which can be expressed as

$$\mu(s, a) = \mathbb{E}\left[r(s, a)\right] = \mathbb{E}_X\left[f(s, a, X)\right]. \qquad (7)$$

For ergodic MDPs, the optimal mean reward ρ^* is independent of the initial state (see [23], section 8.3.3). It is specified as

$$\rho^* = \max_{\beta \in \mathcal{B}} \rho(\beta, \mathbf{M}), \qquad (8)$$

where $\mathcal{B}$ is the set of all policies, $\mathbf{M}$ is the matrix whose (s, a)-th entry is $\mu(s, a)$, and $\rho(\beta, \mathbf{M})$ is the average expected reward per slot using policy β. We use the optimal mean reward as the benchmark and define the cumulative regret of a learning algorithm after T time-slots as

$$\mathfrak{R}(T) := T\rho^* - \mathbb{E}\left[\sum_{t=0}^{T-1} r_t\right]. \qquad (9)$$

This definition of regret of an online learning algorithm is used in reinforcement learning literature [19], [20], [21]. With this definition, the optimal policy also incurs a regret, when the initial state distribution is not same as the stationary distribution. We characterize this regret using tools from Markov chain mixing in appendix B.

B. Optimal Stationary Policy

When the expected rewards for all state-action pairs $\mu(s, a)$ and the transition probabilities $P(s' \mid s, a)$ are known, the problem of determining the optimal policy to maximize the average expected reward over time can be formulated as a linear program (LP) (see e.g. [22], section V.3) shown below.

$$
\begin{aligned}
\max \quad & \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} \pi(s, a) \mu(s, a) \\
\text{s.t.} \quad & \pi(s, a) \geq 0, \ \forall s \in \mathcal{S}, a \in \mathcal{A}_s, \\
& \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} \pi(s, a) = 1, \\
& \sum_{a \in \mathcal{A}_{s'}} \pi(s', a) = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} \pi(s, a) P(s' \mid s, a), \ \forall s' \in \mathcal{S},
\end{aligned}
$$
(10)

where $\pi(s, a)$ denotes the stationary distribution of the MDP. The objective function of the LP from equation (10) gives the average rate corresponding to the stationary distribution $\pi(s, a)$, while the constraints make sure that this stationary distribution corresponds to a valid policy on the MDP. Such LPs can be solved by standard solvers such as CVXPY [24].

If $\pi^*(s, a)$ is the solution to the LP from (10), then for every $s \in \mathcal{S}$, $\pi(s, a) > 0$ for only one action $a \in \mathcal{A}_s$. This is due to the fact the the optimal policy β^* is deterministic for ergodic MDPs in average reward maximization problems (see [23], section 8.3.3). Thus for this problem, $\beta^*(s) = \arg\max_{a \in \mathcal{A}_s} \pi^*(s, a)$. Note that we, henceforth, drop the action index from the stationary distribution, since the policies under consideration are deterministic and the corresponding action is, therefore, deterministically known. In general, we use $\pi_\beta(s)$ to denote the stationary distribution corresponding to the policy β. It must be noted that the stationary distribution of any policy is independent of the reward values and only depends on the transition probability for every state-action pair. The expected average reward depends on the stationary distribution as

$$
\rho(\beta, \mathbf{M}) = \sum_{s \in \mathcal{S}} \pi_\beta(s) \mu(s, \beta(s)).
$$
(11)

In terms of this notation, the LP from (10) equivalent to maximizing $\rho(\beta, \mathbf{M})$ over $\beta \in \mathcal{B}$. Since the matrix $\mathbf{M}$ is unknown, we develop online learning policies for our problem in the next section.

IV. ONLINE LEARNING ALGORITHMS

For the power allocation problem under consideration, although the agent knows the state transition probabilities, the mean rewards for the state-action pairs $\mu(s, a)$ values are still unknown. Hence, the agent cannot solve the LP from (10) to figure out the optimal policy. Any online learning algorithm needs to learn the reward values over time and update its policy adaptively. One interesting aspect of the problem, however, is that the reward function from equation (3) is known to the agent. Since the reward functions under consideration (1) is bijective, once the reward is revealed to the agent, it can infer the instantaneous realization of the random variable X.

This can be used to predict the rewards that would have been obtained at that time for other state-action pairs using the function knowledge.

In our online learning framework, we store the average values of these inferred rewards $\theta(s, a)$ for all state-action pairs. The idea behind our algorithms is to use the estimated sample mean values for the optimization problem instead of the unknown $\mu(s, a)$ values in the objective function of the LP from (10). Since the $\theta(s, a)$ values get updated after each reward revelation, the agent needs to solve the LP again and again. We propose two online learning algorithms: LPSM (linear program of sample means) where the agent solves the LP at each slot and Epoch-LPSM where the LP is solved at fixed pre-defined time slots. Although the agent is unaware of the actual $\mu(s, a)$ values, it learns the statistics $\theta(s, a)$ over time and eventually figures out the optimal policy.

Let $B(s, a) \geq \sup_{x \in \mathcal{X}} f(s, a, x) - \inf_{x \in \mathcal{X}} f(s, a, x)$ denote any upper bound on the maximum possible range of the reward for the state-action pair (s, a) over the support $\mathcal{X}$ of the random variable X. We use following notations in the analysis of our algorithms: $B_0 := \max_{(s, a)} B(s, a)$, $\Delta_1 := \rho^* - \max_{\beta \neq \beta^*} \rho(\beta, \mathbf{M})$. The total number of states and actions are specified as $S := |\mathcal{S}|$, $A := |\mathcal{A}|$, respectively. Also, Θ_t denotes the matrix containing the entries $\theta_t(s, a)$ at time t.

A. LPSM

The LPSM algorithm presented in algorithm 1 solves the LP at each time-step and updates its policy based on the solution obtained. It stores only one θ value per state-action pair. Its required storage is, therefore, $O(SA)$. In theorem 1, we derive an upper bound on the expected number of slots where the LP fails to find the optimal solution during the execution of LPSM. We use this result to bound the total expected regret of LPSM in theorem 2. These results guarantee that the regret is always upper bounded by a constant. Note that, for the ease of exposition, we assume that the time starts at $t = 0$. This simplifies the analysis and has no impact on the regret bounds.

Theorem 1. *The expected number of slots where non-optimal policies are played by LPSM is upper bounded by*

$$
1 + \frac{(1 + A) S}{e^{\frac{1}{2} \left(\frac{\Delta_1}{B_0} \right)^2} - 1}.
$$
(12)

Proof: Let β_t denote the policy obtained by LPSM at time t and $\mathbb{I}(z)$ be the indicator function defined to be 1 when the predicate z is true, and 0 otherwise. Now the number of slots where non-optimal policies are played can be expressed as

$$
\begin{aligned}
N_1 &= 1 + \sum_{t=1}^{\infty} \mathbb{I}\{\beta_t \neq \beta^*\} \\
&= 1 + \sum_{t=1}^{\infty} \mathbb{I}\{\rho(\beta^*, \Theta_t) \leq \rho(\beta_t, \Theta_t)\}.
\end{aligned}
$$
(13)

We observe that $\rho(\beta^*, \Theta_t) \leq \rho(\beta_t, \Theta_t)$ implies that at least one of the following inequalities must be true:

$$
\rho(\beta^*, \Theta_t) \leq \rho(\beta^*, \mathbf{M}) - \frac{\Delta_1}{2}
$$
(14)

6

Algorithm 1 LPSM

1: **Initialization:** For all (s, a) pairs, $\theta(s, a) = 0$.
2: **for** $n = 0$ **do**
3:      Given the state s_0 and choose any valid action;
4:      Update all (s, a) pairs: $\theta(s, a) = f(s, a, x_0)$;
5: **end for**
6: // MAIN LOOP
7: **while** 1 **do**
8:      $n = n + 1$;
9:      Solve the LP from (10) using $\theta(s, a)$ in place of unknown $\mu(s, a)$;
10:      In terms of the LP solution $\pi_{(n)}$, define $\beta_n(s) = \arg\max\limits_{a \in \mathcal{A}_s} \pi_{(n)}(s, a), \forall s \in \mathcal{S}$;
11:      Given the state s_n, select the action $\beta_n(s_n)$;
12:      Update for all valid (s, a) pairs:

$$\theta(s, a) \leftarrow \frac{n\theta(s, a) + f(s, a, x_n)}{n + 1};$$

13: **end while**

$$\rho(\beta_t, \Theta_t) \geq \rho(\beta_t, \mathbf{M}) + \frac{\Delta_1}{2} \tag{15}$$

$$\rho(\beta^*, \mathbf{M}) < \rho(\beta_t, \mathbf{M}) + \Delta_1. \tag{16}$$

Note that the event from condition (16) can never occur, because of the definition of Δ_1. Hence we upper bound the probabilities of the other two events. For the first event from condition (14), we get

$$\Pr\left\{\rho(\beta^*, \Theta_t) \leq \rho(\beta^*, \mathbf{M}) - \frac{\Delta_1}{2}\right\}$$

$$= \Pr\left\{\sum_{s \in \mathcal{S}} \pi^*(s, \beta^*(s))\theta_t(s, \beta^*(s)) \leq \right.$$

$$\left. \sum_{s \in \mathcal{S}} \pi^*(s, \beta^*(s))\mu(s, \beta^*(s)) - \frac{\Delta_1}{2}\right\}$$

$$\leq \Pr\left\{\text{For at least one state } s \in \mathcal{S}: \right.$$

$$\pi^*(s, \beta^*(s))\theta_t(s, \beta^*(s)) \leq$$

$$\left. \pi^*(s, \beta^*(s))\left(\mu(s, \beta^*(s)) - \frac{\Delta_1}{2}\right)\right\}$$

$$\leq \sum_{s \in \mathcal{S}} \Pr\left\{\pi^*(s, \beta^*(s))\theta_t(s, \beta^*(s)) \leq \right.$$

$$\left. \pi^*(s, \beta^*(s))\left(\mu(s, \beta^*(s)) - \frac{\Delta_1}{2}\right)\right\}$$

$$\leq \sum_{s \in \mathcal{S}} \Pr\left\{\theta_t(s, \beta^*(s)) \leq \mu(s, \beta^*(s)) - \frac{\Delta_1}{2}\right\}$$

$$\stackrel{(a)}{\leq} \sum_{s \in \mathcal{S}} e^{-2\left(\frac{\Delta_1}{2B(s, \beta^*(s))}\right)^2 t}$$

$$= S e^{-2\left(\frac{\Delta_1}{2B_0}\right)^2 t}, \tag{17}$$

where (a) holds due to Hoeffding's inequality from lemma 1 (see appendix A).

Similarly for the second event from condition (15), we get

$$\Pr\left\{\rho(\beta_t, \Theta_t) \geq \rho(\beta_t, \mathbf{M}) + \frac{\Delta_1}{2}\right\}$$

$$= \Pr\left\{\sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} \pi_{\beta_t}(s, a)\theta_t(s, a) \geq \right.$$

$$\left. \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} \pi_{\beta_t}(s, a)\mu(s, a) + \frac{\Delta_1}{2}\right\}$$

$$\leq \Pr\left\{\text{For at least one state-action pair } (s, a): \right.$$

$$\left. \pi_{\beta_t}(s, a)\theta_t(s, a) \geq \pi_{\beta_t}(s, a)\left(\mu(s, a) + \frac{\Delta_1}{2}\right)\right\}$$

$$\leq \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} \Pr\left\{\pi_{\beta_t}(s, a)\theta_t(s, a) \geq \right.$$

$$\left. \pi_{\beta_t}(s, a)\left(\mu(s, a) + \frac{\Delta_1}{2}\right)\right\}$$

$$= \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} \Pr\left\{\theta_t(s, a) \geq \mu(s, a) + \frac{\Delta_1}{2}\right\}$$

$$\stackrel{(b)}{\leq} \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} e^{-2\left(\frac{\Delta_1}{2B(s, a)}\right)^2 t}$$

$$\leq S A e^{-2\left(\frac{\Delta_1}{2B_0}\right)^2 t}, \tag{18}$$

where (b) holds due to Hoeffding's inequality from lemma 1 in appendix A.

The expected number of non-optimal policies from equation (13), therefore, can be expressed as

$$\mathbb{E}[N_1] \leq 1 + \sum_{t=1}^{\infty} \Pr\left\{\rho(\beta^*, \Theta_t) \leq \rho(\beta_t, \Theta_t)\right\}$$

$$\leq 1 + \sum_{t=1}^{\infty} \left(\Pr\left\{\rho(\beta^*, \Theta_t) \leq \rho(\beta^*, \mathbf{M}) - \frac{\Delta_1}{2}\right\}\right.$$

$$\left. + \Pr\left\{\rho(\beta_t, \Theta_t) \geq \rho(\beta_t, \mathbf{M}) + \frac{\Delta_1}{2}\right\}\right)$$

$$\leq 1 + (1 + A) S \sum_{t=1}^{\infty} e^{-2\left(\frac{\Delta_1}{2B_0}\right)^2 t}$$

$$\leq 1 + (1 + A) S \frac{e^{-\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2}}{1 - e^{-\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2}}$$

$$\leq 1 + \frac{(1 + A) S}{e^{\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2} - 1}. \tag{19}$$

              ■

It is important to note that even if the optimal policy is found by the LP and played during certain slots, it does not mean that regret contribution of those slots is zero. According to the definition of regret from equation (9), regret contribution of a certain slot is zero if and only if the optimal policy is played and the corresponding Markov chain is at its stationary distribution. In appendix B, we introduce tools to analyze the mixing of Markov chains and characterize this regret contribution in theorem 10. These results are used to upper bound the LPSM regret in the next theorem.

Theorem 2. *The total expected regret of the LPSM algorithm is upper bounded by*

$$\left(1 + \frac{(1+A)\,S}{\mathrm{e}^{\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2} - 1}\right)\left(\frac{\mu_{\max}}{1-\gamma} + \Delta_{\max}\right). \qquad (20)$$

where $\gamma = \max_{s,s' \in \mathcal{S}} \|P_*(s',\cdot) - P_*(s,\cdot)\|_{\mathrm{TV}}$, P_* *denotes the transition probability matrix corresponding to the optimal policy,* $\mu_{\max} = \max_{s \in \mathcal{S}, a \in \mathcal{A}_s} \mu(s,a)$ *and* $\Delta_{\max} = \rho^* - \min_{s \in \mathcal{S}, a \in \mathcal{A}_s} \mu(s,a)$.

Proof: The regret of LPSM arises when either non-optimal actions are taken or when optimal actions are taken and the corresponding Markov chain is not at stationarity. For the first source of regret, it is sufficient to analyze the number of instances where the LP fails to find the optimal policy. For the second source, however, we need to analyze the total number of phases where the optimal policy is found in succession.

Since only the optimal policy is played in consecutive slots in a phase, it corresponds to transitions on the Markov chain associated with the optimal policy and the tools from appendix B can be applied. According to theorem 10, the regret contribution of any phase is bounded from above by $(1-\gamma)^{-1}\mu_{\max}$. As proved in theorem 1, for $t \geq 1$, the expected number of instances of non-optimal policies is upper bounded by $\frac{(1+A)S}{\mathrm{e}^{\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2} - 1}$. Since any two optimal phases must be separated by at least one non-optimal slot, the expected number of optimal phases is upper bounded by $1 + \frac{(1+A)S}{\mathrm{e}^{\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2} - 1}$. Hence, for $t \geq 1$, the expected regret contribution from the slots following the optimal policy is upper bounded by

$$\left(1 + \frac{(1+A)\,S}{\mathrm{e}^{\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2} - 1}\right)\frac{\mu_{\max}}{1-\gamma}. \qquad (21)$$

Note that the maximum regret possible during one slot is $\Delta_{\max}$. Hence for the slots where non-optimal policies are played, the corresponding expected regret contribution is upper bounded by $\left(1 + \frac{(1+A)S}{\mathrm{e}^{\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2} - 1}\right)$.

Overall expected regret for the LPSM algorithm is, therefore, bounded from above by equation (20). ∎

Remark 1. *It must be noted that we call two policies as same if and only if they recommend identical actions for every state. It is, therefore, possible for a non-optimal policy to recommend optimal actions for some of the states. In the analysis of LPSM, we count all occurrences of non-optimal policies as regret contributing occurrences in order to upper bound the regret.*

Remark 2. *Note that the LPSM algorithm presented above works for general reward functions* f. *The rate function in the energy harvesting communications is, however, not dependent on the state which is the battery status. It is a function of the transmit power level and the channel gain only. The LPSM algorithm, therefore, needs to store only one* θ *variable for each transmit power level and needs* $O(A)$ *storage overall. The probability of event from condition (15) is bounded by a*

tighter upper bound of $A\mathrm{e}^{-2\left(\frac{\Delta_1}{2B_0}\right)^2 t}$. *The regret upper bound from theorem 2 is also tightened to*

$$\left(1 + \frac{(S+A)}{\mathrm{e}^{\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2} - 1}\right)\left(\frac{\mu_{\max}}{1-\gamma} + \Delta_{\max}\right). \qquad (22)$$

For the LPSM algorithm, we can prove a stronger result about the convergence time. Let Z be the random variable corresponding to the first time-slot after which LPSM never fails to find the optimal policy. This means that $Z - 1$ represents the last time-slot where LPSM finds a non-optimal policy. We prove that the expected value of Z is finite, which means that LPSM takes only a finite amount time in expectation before it starts following the genie. We present this result in theorem 3.

Theorem 3. *For the LPSM algorithm, the expected value of the convergence time* Z *is finite.*

Proof: Since $Z - 1$ denotes the index of the last slot where LPSM errs, all the slots from Z onward must have found the optimal policy β^*. We use this idea to bound the following probability

$$\Pr\{Z \leq Z_0\}$$
$$= \Pr\{\text{LPSM finds the optimal policy}$$
$$\text{in all slots } Z_0, Z_0 + 1, \cdots\}$$
$$= 1 - \Pr\{\text{LPSM fails in at least one slot}$$
$$\text{in } \{Z_0, Z_0 + 1, \cdots\}\}$$
$$\geq 1 - \sum_{t=Z_0}^{\infty} \Pr\{\text{LPSM fails at } t\}$$
$$\geq 1 - \sum_{t=Z_0}^{\infty}(1+A)\,S\mathrm{e}^{-2\left(\frac{\Delta_1}{2B_0}\right)^2 t} \quad \text{(From equation (19))}$$
$$= 1 - \frac{(1+A)S\mathrm{e}^{-\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2 Z_0}}{1 - \mathrm{e}^{-\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2}}. \qquad (23)$$

We, therefore, get the following exponential inequality for $Z_0 \geq 1$

$$\Pr\{Z > Z_0\} \leq \frac{(1+A)S\mathrm{e}^{-\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2 Z_0}}{1 - \mathrm{e}^{-\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2}}. \qquad (24)$$

The expectation of Z can, now, be bounded as

$$\mathbb{E}[Z] = \sum_{Z_0=0}^{\infty} Z_0 \Pr\{Z = Z_0\}$$
$$= \sum_{Z_0=0}^{\infty} \Pr\{Z > Z_0\}$$
$$\leq 1 + \sum_{Z_0=1}^{\infty} \frac{(1+A)S\mathrm{e}^{-\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2 Z_0}}{1 - \mathrm{e}^{-\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2}}$$
$$= 1 + \frac{(1+A)S\mathrm{e}^{-\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2}}{\left(1 - \mathrm{e}^{-\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2}\right)^2}. \qquad (25)$$

The expected value of Z is, therefore, finite. ∎

Remark 3. *Note that the result about finite expected convergence time is not directly implied by the constant regret result. The proof of theorem 3 relies on the exponential nature of the concentration bound, whereas it is possible to prove constant expected regrets even for weaker concentration bounds [25].*

B. Epoch-LPSM

The main drawback of the LPSM algorithm is that it is computationally heavy as it solves one LP per time-slot. In order to reduce the computation requirements, we propose the Epoch-LPSM algorithm in algorithm 2. Epoch-LPSM solves the LP in each of the first n_0 slots, divides the later time into several epochs and solves the LPs only at the beginning of each epoch. The policy obtained by solving the LP at the beginning of an epoch is followed for the remaining slots in that epoch. We increase the length of these epochs exponentially as time progresses and our confidence on the obtained policy increases. In spite of solving much fewer number of LPs, the regret of Epoch-LPSM is still bounded by a constant. First, we obtain an upper bound on the number of slots where the algorithm plays non-optimal policies in theorem 4 and later use this result to bound the regret in theorem 5.

Algorithm 2 Epoch-LPSM

1: **Parameters:** $n_0 \in \mathbb{N}$ and $\eta \in \{2, 3, \cdots\}$.
2: **Initialization:** $k = 0$, $n = 0$ and for all (s, a) pairs, $\theta(s, a) = 0$.
3: **while** $n < n_0$ **do**
4:    Follow LPSM algorithm to decide action a_n, update the θ variables accordingly and increment n;
5: **end while**
6: **while** $n \geq n_0$ **do**
7:    $n = n + 1$;
8:    **if** $n = n_0 \eta^k$ **then**
9:      $k = k + 1$;
10:     Solve the LP from (10) with $\theta(s, a)$ in place of unknown $\mu(s, a)$;
11:     In terms of the LP solution $\pi_{(n)}$, define $\beta_{(k)}(s) = \arg\max_{a \in \mathcal{A}_s} \pi_{(n)}(s, a)$, $\forall s \in \mathcal{S}$;
12:   **end if**
13:   Given the state s_n, select the action $\beta_{(k)}(s_n)$;
14:   Update for all (s, a) pairs:
$$\theta(s, a) \leftarrow \frac{n\theta(s, a) + f(s, a, x_n)}{n + 1};$$
15: **end while**

Theorem 4. *The expected number of slots where non-optimal policies are played by Epoch-LPSM is upper bounded by*

$$1 + (1+A)S\left(\frac{1 - e^{-\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2 n_0}}{e^{\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2} - 1}\right) + (\eta - 1)(1+A)Sn_0\sigma_{n_0,\eta},$$
(26)

where $\sigma_{n_0,\eta} = \sum_{k=0}^{\infty} \eta^k e^{-\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2 n_0 \eta^k} < \infty$.

Proof: Note that epoch k starts at $t = n_0\eta^{k-1}$ and end at $t = n_0\eta^k - 1$. The policy obtained at $t = n_0\eta^{k-1}$ by solving the LP is, therefore, played for $(\eta^k - \eta^{k-1})n_0$ number of slots.

Let us analyse the probability that the policy played during epoch k is not optimal. Let that policy be $\beta_{(k)}$.

$$\Pr\{\beta_{(k)} \neq \beta^*\} = \Pr\{\beta_{n_0\eta^{k-1}} \neq \beta^*\}$$
$$\overset{(a)}{\leq} (1+A)Se^{-2\left(\frac{\Delta_1}{2B_0}\right)^2 n_0\eta^{k-1}},$$
(27)

where (a) holds for all $k \geq 1$ as shown in the proof of theorem 1.

When the LP fails to obtain the optimal policy at the beginning of an epoch, then all the slots in that epoch will play the obtained non-optimal policy. Let N_2 denote total number of such slots. We get

$$N_2 = 1 + \sum_{t=1}^{n_0-1} \mathbb{I}\{\beta_t \neq \beta^*\}$$
$$+ \sum_{k=1}^{\infty} n_0(\eta^k - \eta^{k-1})\mathbb{I}\{\beta_{(k)} \neq \beta^*\}.$$

In expectation, we get

$$\mathbb{E}[N_2] = 1 + \sum_{t=1}^{n_0-1} \Pr\{\beta_t \neq \beta^*\}$$
$$+ \sum_{k=1}^{\infty} n_0(\eta^k - \eta^{k-1})\Pr\{\beta_{(k)} \neq \beta^*\}$$
$$\leq 1 + \sum_{t=1}^{n_0-1} (1+A)Se^{-2\left(\frac{\Delta_1}{2B_0}\right)^2 t}$$
$$+ \sum_{k=1}^{\infty} n_0(\eta^k - \eta^{k-1})(1+A)Se^{-2\left(\frac{\Delta_1}{2B_0}\right)^2 n_0\eta^{k-1}}$$
$$\leq 1 + (1+A)S\sum_{t=1}^{n_0-1} e^{-\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2 t}$$
$$+ n_0(\eta-1)(1+A)S\sum_{k=0}^{\infty} \eta^k e^{-\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2 n_0\eta^k}$$
$$\leq 1 + (1+A)Se^{-\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2} \frac{1 - e^{-\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2 n_0}}{1 - e^{-\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2}}$$
$$+ n_0(\eta-1)(1+A)S\sigma_{n_0,\eta},$$
(28)

where $\sigma_{n_0,\eta} < \infty$ holds due to ratio test for series convergence as

$$\lim_{k \to \infty} \left| \frac{\eta^{k+1} e^{-\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2 n_0\eta^{k+1}}}{\eta^k e^{-\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2 n_0\eta^k}} \right|$$
$$= \lim_{k \to \infty} \left| \eta e^{-\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2 n_0(\eta^{k+1} - \eta^k)} \right|$$
$$= \lim_{k \to \infty} \left| \eta \left(e^{\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2 n_0(\eta-1)} \right)^{-\eta^k} \right|$$
$$= 0.$$

Now we analyse the regret of Epoch-LPSM in the following theorem.

Theorem 5. *The total expected regret of the Epoch-LPSM algorithm is upper bounded by*

$$
\left(1 + (1+A)S\left(\frac{1 - \mathrm{e}^{-\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2 n_0}}{\mathrm{e}^{\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2} - 1}\right)\right)\left(\frac{\mu_{\max}}{1-\gamma} + \Delta_{\max}\right)
$$
$$
+ (\eta - 1)(1+A)Sn_0\sigma_{n_0,\eta}\Delta_{\max}
$$
$$
+ (1+A)S\left(\sum_{k=1}^{\infty}\mathrm{e}^{-2\left(\frac{\Delta_1}{2B_0}\right)^2 n_0\eta^{k-1}}\right)\frac{\mu_{\max}}{1-\gamma}. \quad (29)
$$

Proof: First we analyse the regret contribution from the first n_0 slots. As argued in the proof of theorem 2, the regret contribution of the first n_0 slots is upper bounded by

$$
\left(1 + (1+A)S\left(\frac{1 - \mathrm{e}^{-\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2 n_0}}{\mathrm{e}^{\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2} - 1}\right)\right)\left(\frac{\mu_{\max}}{1-\gamma} + \Delta_{\max}\right). \quad (30)
$$

Now we analyse the number of phases where the optimal policy is played in successive slots for $t \geq n_0$. Note that any two optimal phases are separated by at least one non-optimal epoch. We bound the number of non-optimal epochs N_3 as

$$
\mathbb{E}[N_3] = \sum_{k=1}^{\infty}\Pr\{\beta_{(k)} \neq \beta^*\}
$$
$$
\leq (1+A)S\sum_{k=1}^{\infty}\mathrm{e}^{-2\left(\frac{\Delta_1}{2B_0}\right)^2 n_0\eta^{k-1}}
$$
$$
\text{(From equation (27))}
$$
$$
< 0,
$$

where the series $\sum_{k=1}^{\infty}\mathrm{e}^{-2\left(\frac{\Delta_1}{2B_0}\right)^2 n_0\eta^{k-1}}$ converges due to ratio test as

$$
\lim_{k\to\infty}\left|\frac{\mathrm{e}^{-\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2 n_0\eta^{k+1}}}{\mathrm{e}^{-\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2 n_0\eta^{k}}}\right|
$$
$$
= \lim_{k\to\infty}\left|\left(\mathrm{e}^{\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2 n_0(\eta-1)}\right)^{-\eta^k}\right| = 0.
$$

Hence for $t \geq n_0$, there can be at most $\mathbb{E}[N_3]$ number of optimal phases in expectation. Since each of these phases can contribute a maximum of $(1-\gamma)^{-1}\mu_{\max}$ regret in expectation, total regret from slots with optimal policies for $t \geq n_0$ is upper bounded by $\mathbb{E}[N_3]\frac{\mu_{\max}}{1-\gamma}$. Also, the expected number of slots where a non-optimal policy is played for $t \geq n_0$ is bounded by $(\eta-1)(1+A)Sn_0\sigma_{n_0,\eta}$ as derived in the proof of theorem 4. The regret contribution of these slots is, therefore, bounded by $(\eta-1)(1+A)Sn_0\sigma_{n_0,\eta}\Delta_{\max}$, since the maximum expected regret incurred during any slot is $\Delta_{\max}$.

The total expected regret of Epoch-LPSM is, therefore, upper bounded by the expression (29). ∎

Remark 4. *Similar to the LPSM algorithm, the algorithm presented above works for general reward functions f. Since the rate function in the energy harvesting communications is*

not dependent on the state, the Epoch-LPSM algorithm needs to store only one θ variable per transmit power level and uses $O(A)$ storage overall. The regret upper bound from theorem 5 is also tightened to

$$
\left(1 + (S+A)\left(\frac{1 - \mathrm{e}^{-\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2 n_0}}{\mathrm{e}^{\frac{1}{2}\left(\frac{\Delta_1}{B_0}\right)^2} - 1}\right)\right)\left(\frac{\mu_{\max}}{1-\gamma} + \Delta_{\max}\right)
$$
$$
+ (\eta - 1)(S+A)n_0\sigma_{n_0,\eta}\Delta_{\max}
$$
$$
+ (S+A)\left(\sum_{k=1}^{\infty}\mathrm{e}^{-2\left(\frac{\Delta_1}{2B_0}\right)^2 n_0\eta^{k-1}}\right)\frac{\mu_{\max}}{1-\gamma}. \quad (31)
$$

C. Regret vs Computation Tradeoff

The LPSM algorithm solves T LPs in time T, whereas the Epoch-LPSM algorithm solves n_0 LPs in the initial n_0 slots and $\lceil\log_\eta(T - n_0)\rceil$ LPs when the time gets divided into epochs. This drastic reduction in the required computation comes at the cost of an increase in the regret for Epoch-LPSM. It must, however, be noted that both the algorithms have constant-bounded regrets. Also, increasing the value of the parameter η in Epoch-LPSM leads to reduction in the number of LPs to be solved over time by increasing the epoch lengths. Any non-optimal policy found by LP, therefore, gets played over longer epochs increasing the overall regret. Increasing n_0 increases the total number of LPs solved by the algorithm while reducing the expected regret. The system designer can analyse the regret bounds of these two algorithms and its own performance requirements to choose the parameters n_0 and η for the system. We analyse the effect of variation of these parameters on the regret performance of Epoch-LPSM through numerical simulations in section VII.

V. MULTI-CHANNEL COMMUNICATION

In this section, we extend the energy harvesting communications problem to consider a system where there exists a set of parallel channels, with unknown statistics, for communication and one of these channels is to be selected in each slot. The goal is to utilize the battery at the transmitter and maximize the amount of data transmitted over time. Given a time-slotted system, we assume that the agent is aware of the distribution of energy arrival. The agent sees the current state of the battery and needs to decide the transmit power-level and the channel to be used for transmission. This problem, therefore, involves an additional decision making layer compared to the single channel case. Note that we use the terms transmit power and action interchangeably in this section.

For this problem, we simplify the notations used previously and drop the state as a parameter for the reward function f, since the rate is not a function of the battery state and only depends on the transmit power-level and the channel gain. These channels will, in general, have different distributions of channel gains and there may not be a single channel that is optimal for all transmit power-levels. The expected rate achieved by selecting j-th channel from the set of M channels and an action a corresponding the transmit power used is

$$
\mu_j(a) = \mathbb{E}_{X_j}[f(a, X_j)], \quad (32)
$$

where X_j denotes the random gain of j-th channel. Let us define $\phi^* : \mathcal{A} \to \{1, 2 \cdots, M\}$ as the mapping from transmit power-levels to their corresponding optimal channels:

$$\phi^*(a) = \arg\max_{j \in \{1, 2 \cdots, M\}} \mu_j(a). \qquad (33)$$

A genie that knows the distributions of different channels gains can figure out the optimal channel mapping ϕ^*. Once an action a is chosen by the genie, there is no incentive to use any channel other than $\phi^*(a)$ for transmission during that slot. Let $\mu^*(a)$ denote the expected rate of the best channel for action a, i.e. $\mu^*(a) = \mu_{\phi^*(a)}(a)$. The genie uses these values to solve the following LP.

$$\max \quad \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} \pi(s, a) \mu^*(a)$$

$$\text{s.t.} \quad \pi(s, a) \geq 0, \ \forall s \in \mathcal{S}, a \in \mathcal{A}_s,$$

$$\sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} \pi(s, a) = 1,$$

$$\sum_{a \in \mathcal{A}_{s'}} \pi(s', a) = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} \pi(s, a) P(s' \mid s, a), \ \forall s' \in \mathcal{S}.$$

$$(34)$$

The genie obtains $\beta^* : \mathcal{S} \to \mathcal{A}$, the optimal mapping from the battery state to the transmit power-level using the non-zero terms of the optimal stationary distribution $\pi^*(s, a)$. Note that the constraints of the optimization problem (34) ensure that the stationary distribution actually corresponds to some valid deterministic state-action mapping.

Let $\mathcal{B}$ be the set of all state-action mappings. There are only a finite number of such mappings $\beta \in \mathcal{B}$ and the stationary distribution only depends on the matrix of state transition probabilities which is assumed to be known. We use $\pi_\beta(s)$ denote the stationary distribution corresponding to the state-action mapping β. We dropped the action parameter from the previous notation, since it is implicit from β. The expected average reward of the power selection policy β along with a channel selection policy ϕ is calculated as

$$\rho(\beta, \phi, \mathbf{M}) = \sum_{s \in \mathcal{S}} \pi_\beta(s) \mu_{\phi(a)}(a), \qquad (35)$$

where $\mathbf{M}$ denotes the matrix containing all $\mu_j(a)$ values for power-channel pairs. For the genie under consideration, the LP from (34) is equivalent to

$$\beta^* = \arg\max_{\beta \in \mathcal{B}} \rho(\beta, \phi^*, \mathbf{M}). \qquad (36)$$

The expected average reward of the genie can, therefore, be defined as $\rho^* = \rho(\beta^*, \phi^*, \mathbf{M})$. Since the mean rate matrix $\mathbf{M}$ is unknown to the agent, we propose an online learning framework for this problem.

A. Online Learning Algorithm

Since the agent does not know the distributions of channel gains, it needs to learn the rates for various power-channel pairs, figure out ϕ^* over time and use it to make decisions about the transmit power-level at each slot. We propose an online learning algorithm called Multi-Channel LPSM (MC-LPSM) for this problem. We analyze the performance of MC-LPSM in terms of the regret as defined in equation (9).

The MC-LPSM algorithm stores estimates of the rates for all power-level and channel pairs based on the observed values of the channel gains. Whenever the rate obtained is revealed to the agent, it can infer the instantaneous gain of the chosen channel knowing the transmit power-level. Once the instantaneous gain of a channel is known, this information can be used to update the sample-mean rate estimates of all the power-levels for that channel. The algorithm divides time into two interleaved sequences: an exploration sequence and an exploitation sequence similar to the DSEE algorithm from [9]. In the exploitation sequence, the agent uses its current estimates of the rates to determine the transmit power and channel selection policies. First it selects a channel for each power-level that has the highest empirical rate for that transmit power. The sample-mean rate estimates for these power-channel pairs are, then, used to solve the LP from equation (34) with θ values replacing the μ values and to obtain a power-selection policy for that slot. In the exploration sequence, the agent selects all channels in a round-robin fashion in order to learn the rates over time and chooses the transmit power-levels arbitrarily. The choice of the length of the exploration sequence balances the tradeoff between exploration and exploitation.

Let $\mathcal{R}(t)$ denote the set of time indexes that are marked as exploration slots up to time t. Let $|\mathcal{R}(t)|$ be the cardinality of the set $\mathcal{R}(t)$. At any given time t, m_j stores the number of times j-th channel has been chosen during the exploration sequence till that slot. Using these notations we present MC-LPSM in algorithm 3. Note that MC-LPSM stores a $\theta_j(a)$ variable for every action action-channel pair (a, j) and an m_j variable for every channel j. It, therefore, requires $O(MA)$ storage.

Note that the agent updates the θ variables only during the exploration sequence when it tries different channels sequentially. Since the θ variables do not change during exploitation, the agent does not have to solve the LP in all exploitation slots. During a phase of successive exploitation slots, the channel and power selection policies obtained by solving the LPs remain unchanged. The agent, therefore, needs to solve the LP at time t only if the previous slot was an exploration slot. Since there are at most $|\mathcal{R}(T)|$ exploration slots, MC-LPSM solves at most $|\mathcal{R}(T)|$ number of LPs in T slots.

B. Regret Analysis of MC-LPSM

Let us first define the notations used in the regret analysis. Since the agent is unaware of the matrix of expected rates $\mathbf{M}$, it stores the estimates of the expected rates in matrix Θ. We define $\rho(\beta, \phi, \Theta)$ according to equation (35) with the actual mean values replaced by their corresponding estimates in Θ. Let P_* denote the transition probability matrix corresponding to the optimal state-action mapping β^*. We further define:

$$\gamma = \max_{s, s' \in \mathcal{S}} \|P_*(s', \cdot) - P_*(s, \cdot)\|_{\mathrm{TV}} \qquad (38)$$

$$\mu_{\max} = \max_{a \in \mathcal{A}} \mu^*(a) \qquad (39)$$

$$\Delta_{\max} = \rho^* - \min_{a \in \mathcal{A}, \ j \in \{1, 2, \cdots, M\}} \mu_j(a) \qquad (40)$$

$$\Delta_3 = \min_{a \in \mathcal{A}, \ j \neq \phi^*(a)} \{\mu^*(a) - \mu_j(a)\} \qquad (41)$$

Algorithm 3 MC-LPSM

1: **Parameters:** $w > \frac{2B_0^2}{d^2}$.

2: **Initialization:** For all $a \in \mathcal{A}$ and $j \in \{1, 2, \cdots, M\}$, $\theta_j(a) = 0$. Also $m_j = 0$ for all channels j and $n = 0$.

3: **while** $n < T$ **do**

4:    $n = n + 1$;

5:    **if** $n \in \mathcal{R}(T)$ **then**

6:       // Exploration sequence

7:       Choose channel $j = ((n-1) \mod M) + 1$ and any valid power-level as action a;

8:       Update $\theta_j(a)$ variables for all actions $a \in \mathcal{A}$ for the chosen channel j:

$$\theta_j(a) \leftarrow \frac{m_j \theta_j(a) + f(a, x_n)}{m_j + 1},$$
$$m_j \leftarrow m_j + 1; \tag{37}$$

9:    **else**

10:       // Exploitation sequence

11:       **if** $n - 1 \in \mathcal{R}(T)$ **then**

12:          Define a channel mapping ϕ, such that $\phi(a) = \max_j \theta_j(a)$;

13:          Solve the LP from (34) with $\theta_{\phi(a)}(a)$ instead of unknown $\mu^*(a)$ for all valid state-action pairs (s, a);

14:          In terms of the LP solution $\pi_{(n)}$, define $\beta(s) = \arg\max_{a \in \mathcal{A}_s} \pi_{(n)}(s, a)$, $\forall s \in \mathcal{S}$;

15:       **end if**

16:       Given the state s_n, select the power-level $a_n = \beta(s_n)$ as action for transmission over channel $\phi(a_n)$;

17:    **end if**

18: **end while**

$$\Delta_4 = \rho^* - \max_{\beta \neq \beta^*} \rho(\beta, \phi^*, \mathbf{M}) \tag{42}$$

$$B_0 = \sup_{x \in \mathcal{X}} f(a, x) - \inf_{x \in \mathcal{X}} f(a, x). \tag{43}$$

In terms of these notations, we provide an upper bound on the regret of MC-LPSM as follows.

Theorem 6. *Given a constant $d \leq \min\{\Delta_3, \Delta_4\}$, choose a constant $w > \frac{2B_0^2}{d^2}$. Construct an exploration sequence sequentially as follows:*

1) *$|\mathcal{R}(1)| < \{1\}$,*

2) *For any $t > 1$, include t in $\mathcal{R}(t)$ iff $|\mathcal{R}(t-1)| < M\lceil w \ln t \rceil$.*

Under this exploration sequence $\mathcal{R}$, the T-slot expected regret of MC-LPSM algorithm is upper bounded by

$$\left(M\lceil w \ln T \rceil + 2AM c_{\left(\frac{wd^2}{2B_0^2}\right)} \right) \left(\Delta_{\max} + \frac{\mu_{\max}}{1 - \gamma} \right), \tag{44}$$

where $c_{(x)} = \sum_{t=1}^{\infty} t^{-x} < \infty$ for $x > 1$.

Proof: In order to upper bound the regret of the MC-LPSM algorithm, we analyze the number of time-slots where the agent plays policy combinations other than (β^*, ϕ^*). Such a failure event at time t corresponds to at least one of the following cases

1) $t \in \mathcal{R}$, i.e. the exploration of different channels,

2) $\phi_t \neq \phi^*$ during exploitation,

3) $\beta_t \neq \beta^*$ during exploitation.

Let $N_4(T)$ be the total number of exploitation slots where MC-LPSM fails to find the optimal power-channel mapping ϕ^* or the optimal state-action mapping β^* up to time T. Let us define events $\mathcal{E}_{1,t} = \{\phi_t \neq \phi^*\}$ and $\mathcal{E}_{2,t} = \{\beta_t \neq \beta^*\}$. Now $N_4(T)$ can be expressed as

$$N_4(T) = \sum_{t \notin \mathcal{R}, t \leq T} \mathbb{I}\{\mathcal{E}_{1,t} \cup \mathcal{E}_{2,t}\}$$
$$= \sum_{t \notin \mathcal{R}, t \leq T} \left(\mathbb{I}\{\mathcal{E}_{1,t}\} + \mathbb{I}\{\mathcal{E}_{2,t} \cap \overline{\mathcal{E}_{1,t}}\} \right). \tag{45}$$

We analyse the two events separately and upper bound their probabilities.

1) Non-Optimal Power-Channel Mapping: We use $\theta_t^*(a)$ to denote $\theta_{\phi^*(a),t}(a)$ for all action a. The probability of the event $\mathcal{E}_{1,t}$ can be bounded as

$$\Pr\{\phi_t \neq \phi^*\} = \Pr \{\text{For at least one action } a \in \mathcal{A} \text{ such that:}$$
$$\phi_t(a) \neq \phi^*(a)\}$$
$$\leq \sum_{a \in \mathcal{A}} \Pr\{\phi_t(a) \neq \phi^*(a)\}$$
$$\leq \sum_{a \in \mathcal{A}} \Pr \{\text{For at least one channel } j \neq \phi^*(a):$$
$$\theta_{j,t}(a) \geq \theta_t^*(a)\}$$
$$= \sum_{a \in \mathcal{A}} \sum_{j \neq \phi^*(a)} \Pr\{\theta_{j,t}(a) \geq \theta_t^*(a)\}. \tag{46}$$

In order for the condition $\theta_{j,t}(a) \geq \theta_t^*(a)$ to hold, at least one of the following must hold:

$$\theta_{j,t}(a) \geq \mu_j(a) + \frac{\Delta_3}{2} \tag{47}$$

$$\theta_t^*(a) \leq \mu^*(a) - \frac{\Delta_3}{2} \tag{48}$$

$$\mu^*(a) < \mu_j(a) + \Delta_3. \tag{49}$$

Note that condition (49) cannot hold due to the definition of Δ_3. Hence we upper bound the the probabilities of the other two events. The construction of the exploration sequence guarantees that at $t \notin \mathcal{R}$ each channel has been explored at least $\lceil w \ln t \rceil$ times. Since B_0 upper bounds the maximum deviation in the range of rate values over channels, we bound the probability for the event from condition (47) using Hoeffding's inequality as

$$\Pr \left\{ \theta_{j,t}(a) \geq \mu_j(a) + \frac{\Delta_3}{2} \right\} \leq e^{-\frac{1}{2}\left(\frac{\Delta_3}{B_0}\right)^2 \lceil w \ln t \rceil}$$
$$\leq e^{-\frac{1}{2}\left(\frac{\Delta_3}{B_0}\right)^2 w \ln t}$$
$$\leq t^{-\frac{w}{2}\left(\frac{\Delta_3}{B_0}\right)^2}. \tag{50}$$

Using the Hoeffding's inequality again for the condition (48), we similarly obtain

$$\Pr \left\{ \theta_t^*(a) \leq \mu^*(a) - \frac{\Delta_3}{2} \right\} \leq t^{-\frac{w}{2}\left(\frac{\Delta_3}{B_0}\right)^2}. \tag{51}$$

We can, therefore, express the upper bound from equation (46) as

$$\Pr\{\phi_t \neq \phi^*\} \leq \sum_{a \in \mathcal{A}} \sum_{j \neq \phi^*(a)} \left(\Pr\left\{ \theta_{j,t}(a) \geq \mu_j(a) + \frac{\Delta_3}{2} \right\} \right.$$
$$\left. + \Pr\left\{ \theta_t^*(a) \leq \mu^*(a) - \frac{\Delta_3}{2} \right\} \right)$$
$$\leq \sum_{a \in \mathcal{A}} \sum_{j \neq \phi^*(a)} 2t^{-\frac{w}{2}(\frac{\Delta_3}{B_0})^2}$$
$$\leq 2A(M-1)t^{-\frac{w}{2}(\frac{\Delta_3}{B_0})^2}. \qquad (52)$$

2) Non-Optimal State-Action Mapping: We analyse the event $\mathcal{E}_{2,t} \cap \overline{\mathcal{E}_{1,t}}$ where the LP fails to find the optimal state-action mapping β^* in spite of having found the optimal power-channel mapping ϕ^*.

$$\Pr\{\mathcal{E}_{2,t} \cap \overline{\mathcal{E}_{1,t}}\} = \Pr\{\beta_t \neq \beta_t; \phi_t = \phi^*\}$$
$$= \Pr\{\rho(\beta^*, \phi^*, \Theta_t) \leq \rho(\beta_t, \phi^*, \Theta_t)\}. \quad (53)$$

For $\rho(\beta^*, \phi^*, \Theta_t) \leq \rho(\beta_t, \phi^*, \Theta_t)$ to hold, at least one of the following must hold:

$$\rho(\beta^*, \phi^*, \Theta_t) \leq \rho(\beta^*, \phi^*, \mathbf{M}) - \frac{\Delta_4}{2} \qquad (54)$$

$$\rho(\beta_t, \phi^*, \Theta_t) \geq \rho(\beta_t, \phi^*, \mathbf{M}) + \frac{\Delta_4}{2} \qquad (55)$$

$$\rho(\beta^*, \phi^*, \mathbf{M}) < \rho(\beta_t, \phi^*, \mathbf{M}) + \Delta_4. \qquad (56)$$

The condition from equation (56) cannot hold due to the definition of Δ_4. We use the techniques from equations (17) and (18) in the proof theorem 1 to upper bound the probabilities of the events of equations (54) and (55) as

$$\Pr\left\{ \rho(\beta^*, \phi^*, \Theta_t) \leq \rho(\beta^*, \phi^*, \mathbf{M}) - \frac{\Delta_4}{2} \right\}$$
$$\leq \min\{S, A\} e^{-\frac{1}{2}(\frac{\Delta_4}{B_0})^2 \lceil w \ln t \rceil}$$
$$\leq A t^{-\frac{w}{2}(\frac{\Delta_4}{B_0})^2} \qquad (57)$$

$$\Pr\left\{ \rho(\beta_t, \phi^*, \Theta_t) \geq \rho(\beta_t, \phi^*, \mathbf{M}) + \frac{\Delta_4}{2} \right\}$$
$$\leq A e^{-\frac{1}{2}(\frac{\Delta_4}{B_0})^2 \lceil w \ln t \rceil}$$
$$\leq A t^{-\frac{w}{2}(\frac{\Delta_4}{B_0})^2}. \qquad (58)$$

Note that these concentration bounds are different from the single channel case, as the number of observations leading to θ_t is only $\lceil w \ln t \rceil$ in contrast to t observations for the single channel. Now we update the upper bound from equation (53) as

$$\Pr\{\mathcal{E}_{2,t} \cap \overline{\mathcal{E}_{1,t}}\} \leq \Pr\left\{ \rho(\beta^*, \phi^*, \Theta_t) \leq \rho(\beta^*, \phi^*, \mathbf{M}) - \frac{\Delta_4}{2} \right\}$$
$$+ \Pr\left\{ \rho(\beta_t, \phi^*, \Theta_t) \geq \rho(\beta_t, \phi^*, \mathbf{M}) + \frac{\Delta_4}{2} \right\}$$
$$\leq 2A t^{-\frac{w}{2}(\frac{\Delta_4}{B_0})^2}. \qquad (59)$$

The expected number of exploitation slots, where non-optimal power and channel selection decisions are made

$\mathbb{E}[N_4(T)]$, can be bounded using equations (52) and (59) as

$$\mathbb{E}[N_4(T)] \leq \sum_{t=1}^{T} \left(\Pr\{\mathcal{E}_{1,t}\} + \Pr\{\mathcal{E}_{2,t} \cap \overline{\mathcal{E}_{1,t}}\} \right)$$
$$\leq \sum_{t=1}^{T} \left(2A(M-1)t^{-\frac{w}{2}(\frac{\Delta_3}{B_0})^2} + 2A t^{-\frac{w}{2}(\frac{\Delta_4}{B_0})^2} \right)$$
$$\leq 2AM \sum_{t=1}^{T} t^{-\frac{w}{2}(\frac{d}{B_0})^2}$$
$$\leq 2AM c_{(\frac{wd^2}{2B_0^2})}, \qquad (60)$$

where $d \leq \min\{\Delta_3, \Delta_4\}$. Since $w > \frac{2B_0^2}{d^2}$, the upper bound from equation (60) holds. The expected number of slots, where non-optimal decisions are made including exploration and exploitation sequences, is upper bounded by $M\lceil w \ln T \rceil + 2AM c_{(\frac{wd^2}{2B_0^2})}$. This implies that there can at most be $M\lceil w \ln T \rceil + 2AM c_{(\frac{wd^2}{2B_0^2})}$ phases where the optimal policies are played in succession, since any two optimal phases must have at least one non-optimal slot in between. The total expected regret of any optimal phase is bounded by $\frac{\mu_{\max}}{1-\gamma}$ and the expected regret incurred during a non-optimal slot is bounded by $\Delta_{\max}$. The total expected T-slot regret of the MC-LPSM algorithm is, therefore, upper bounded by the expression (44). ∎

Note that the length of the exploration sequence specified in theorem 6 scales logarithmically in time. The MC-LPSM algorithm using this exploration sequence, therefore, solves $O(\ln T)$ number of LPs in T slots, similar in order to the single channel Epoch-LPSM algorithm.

It must be noted that the logarithmic order regret is achievable by MC-LPSM if we know d, a lower bound on Δ_3 and Δ_4. This is required in order to define a constant w that leads to the series convergence in the regret proof. If no such knowledge is available, the exploration sequence needs to be expanded in order to achieve a regret that is arbitrarily close to the logarithmic order, similar to the DSEE techniques from [9]. The regret result for such an exploration sequence is as follows:

Theorem 7 (Theorem 2 from [9]). *Let g be any positive, monotonically non-decreasing sequence with $g(t) \to \infty$ as $t \to \infty$. Construct an exploration sequence as follows: for any $t > 1$, include t in $\mathcal{R}$ iff $|\mathcal{R}(t-1)| < M\lceil g(t) \ln t \rceil$. Under this exploration sequence $\mathcal{R}$, the T-slot expected regret of MC-LPSM algorithm is $O(g(T) \ln T)$.*

While this regret is not logarithmic in time, one can approach arbitrarily close to the logarithmic order by reducing the diverging rate of $g(t)$. With this construction of the exploration sequence, the MC-LPSM algorithm solves $O(g(T) \ln T)$ number of LPs in time T.

C. Asymptotic Lower Bound

In the multi-channel scenario, there exist one or more channels that are optimal for some transmit power levels with non-zero stationary probability. For every optimal channel j,

there exists some state $s \in \mathcal{S}$ such that $\phi^*(\beta^*(s)) = j$. There may also exist arms that are either not optimal for any transmit power level or are optimal for power levels that have zero stationary probability. We now present an asymptotic lower-bound on regret of any algorithm for the multi-channel energy harvesting communications problem under certain conditions. To prove the regret bound, we first present a lower bound on the number of plays of the non-optimal channels for any algorithm. Our analysis is based on the asymptotic lower bound on the regret of the standard MAB problem by Lai and Robbins [12]. This MAB regret lower bound applies to the settings where the arm-distributions are characterized by a single parameter. This result was extended by Burnetas and Katehakis to distributions indexed by multiple parameters in [26]. In our analysis, however, we restrict ourselves to the single parameter channel-gain distributions.

Let the gain distribution of each channel be expressed by its density function $g(x; \psi)$ with respect to some measure ν, where the density function $g(\cdot; \cdot)$ is known and ψ is an unknown parameter from some set Ψ. Although we consider continuous distributions here, the analysis also holds for discrete distributions where probability mass function replaces the density and the summations replace the integrals. Corresponding to a valid transmit power a and parameter $\psi \in \Psi$, we define the expected rate as

$$\mu(a; \psi) = \int_{x \in \mathcal{X}} f(a, x) g(x; \psi) d\nu(x). \qquad (61)$$

Let $\mathrm{I}(\psi, \psi')$ denote the Kullback-Leibler distance defined as

$$\mathrm{I}(\psi, \psi') = \int_{x \in \mathcal{X}} \left[\ln \left(\frac{g(x; \psi)}{g(x; \psi')} \right) \right] g(x; \psi) d\nu(x). \qquad (62)$$

We now make following assumptions about the density and the parameter set under consideration.

A1 Existence of mean: $\mu(a; \psi) < \infty$ exists for any $\psi \in \Psi$ and $a \in \mathcal{A}$.

A2 Denseness of Ψ: $\forall \psi \in \Psi$, $\forall a \in \mathcal{A}$ and $\forall \delta > 0$, $\exists \psi' \in \Psi$ such that $\mu(a; \psi) < \mu(a; \psi') < \mu(a; \psi) + \delta$.

A3 Positivity of distance: $0 < \mathrm{I}(\psi, \psi') < \infty$ whenever $\mu(a; \psi) < \mu(a; \psi')$ for some $a \in \mathcal{A}$.

A4 Continuity of $\mathrm{I}(\psi, \psi')$: $\forall \epsilon > 0$, $\forall a \in \mathcal{A}$ and $\forall \psi, \psi' \in \Psi$ such that $\mu(a; \psi) < \mu(a; \psi')$, $\exists \delta = \delta(a, \epsilon, \psi, \psi') > 0$ for which $|\mathrm{I}(\psi, \psi') - \mathrm{I}(\psi, \psi'')| < \epsilon$ whenever $\mu(a; \psi') < \mu(a; \psi'') < \mu(a; \psi') + \delta$.

For channel gain distributions satisfying these conditions, we present a lower bound on the number of plays of a non-optimal arm based on the techniques from [12].

Theorem 8. *Assume that the density and the parameter set satisfy assumptions A1-A4. Let $\boldsymbol{\psi} = (\psi_1, \psi_2, \cdots, \psi_M)$ be a valid parameter vector characterizing the distributions of the M channels, $\mathbb{P}_{\boldsymbol{\psi}}$ and $\mathbb{E}_{\boldsymbol{\psi}}$ be the probability measure and expectation under $\boldsymbol{\psi}$. Let $\mathcal{L}$ be any allocation rule that satisfies for every $\boldsymbol{\psi}$ as $T \to \infty$, $\mathfrak{R}_{\mathcal{L}}(T) = o(T^b)$ for every $b > 0$ over an MDP $\mathcal{M}$. Let $N_i(T)$ denote the number of plays of i-th channel up to time T by the rule $\mathcal{L}$, and $\mathcal{O}_{\boldsymbol{\psi}}$ the index set of the optimal channels under the parameter vector $\boldsymbol{\psi}$. Then for*

every channel $i \in \overline{\mathcal{O}_{\boldsymbol{\psi}}}$,

$$\liminf_{T \to \infty} \mathbb{E}_{\boldsymbol{\psi}} \left[\frac{N_i(T)}{\ln T} \right] \geq \max_{j \in \mathcal{O}_{\boldsymbol{\psi}}} \frac{1}{\mathrm{I}(\psi_i, \psi_j)}. \qquad (63)$$

Proof: Without the loss of generality, we assume that $1 \in \overline{\mathcal{O}_{\boldsymbol{\psi}}}$ and $2 \in \mathcal{O}_{\boldsymbol{\psi}}$ for the parameter vector $\boldsymbol{\psi}$. This means that $\exists a \in \mathcal{A}$ such that $\mu(a; \psi_2) > \mu(a; \psi_1)$ and $\mu(a; \psi_2) \geq \mu(a; \psi_j)$ for $3 \leq j \leq M$. Fix any $0 < \delta < 1$. By assumptions A2 and A4, we can choose $\sigma \in \Psi$ such that

$$\mu(a; \sigma) > \mu(a; \psi_2) \ \& \ |\mathrm{I}(\psi_1, \sigma) - \mathrm{I}(\psi_1, \psi_2)| < \delta \mathrm{I}(\psi_1, \psi_2). \qquad (64)$$

Let us define a new parameter vector $\boldsymbol{\sigma} = (\sigma, \psi_2, \cdots, \psi_M)$ such that under $\boldsymbol{\sigma}$, $1 \in \mathcal{O}_{\boldsymbol{\sigma}}$. The basic argument is that any algorithm incurring regrets of order $o(T^b)$ for every $b > 0$ must play every channel a minimum number of times to be able to distinguish between the cases $\boldsymbol{\psi}$ and $\boldsymbol{\sigma}$.

Let $N_{i,j}(T)$ denote the number of times the i-th channel has been played up to time T with power levels for which j-th channel was the optimal channel. We, therefore, have $N_i(T) = \sum_{j=1}^M N_{i,j}(T)$ where $N_{i,j}(T) \geq 0$ for all (i, j) pairs. We define $T_i(T)$ as the number of plays of the power levels for which i-th channel is optimal up to time T. This implies that the allocation rule $\mathcal{L}$ plays channels other than the i-th channel for $T_i(T) - N_{i,i}(T)$ number of times where they were non-optimal. Fix $0 < b < \delta$. Since $\mathfrak{R}_{\mathcal{L}}(T) = o(T^b)$ as $T \to \infty$, we have $\mathbb{E}_{\boldsymbol{\sigma}}[N_{j,i}(T)] = o(T^b)$ when $i \neq j$. Hence for distributions parametrized by $\boldsymbol{\sigma}$, we have

$$\mathbb{E}_{\boldsymbol{\sigma}}[T_1(T) - N_{1,1}(T)] = \sum_{j \neq 1} \mathbb{E}_{\boldsymbol{\sigma}}[N_{j,1}(T)] = o(T^b). \qquad (65)$$

We define a stationary distribution over channel plays under the optimal power selection policy for the MDP as

$$\pi_j = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}, \ \beta^*(a) = j} \pi^*(s, a). \qquad (66)$$

Note that the optimal policies ϕ^* and β^* are dependent on the choice of the parameter vector characterizing the channels and so is π_j. For channels $j \in \mathcal{O}_{\boldsymbol{\sigma}}$, $\pi_j > 0$ under $\boldsymbol{\sigma}$. Since $\mathfrak{R}_{\mathcal{L}}(T) = o(T^b)$ asymptotically, we have $\mathbb{E}_{\boldsymbol{\sigma}}[|\pi_1 T - T_1(T)|] = o(T^b)$ as $T \to \infty$. Fix $0 < c < 1$. We have

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\sigma}} \{T_1(T) \leq (1 - c)\pi_1 T\} \\
&= \mathbb{P}_{\boldsymbol{\sigma}} \{\pi_1 T - T_1(T) \geq c\pi_1 T\} \\
&\leq \mathbb{P}_{\boldsymbol{\sigma}} \{|\pi_1 T - T_1(T)| \geq c\pi_1 T\} \\
&\leq \frac{\mathbb{E}_{\boldsymbol{\sigma}}[|\pi_1 T - T_1(T)|]}{c\pi_1 T} \qquad \text{(Markov's Inequality)} \\
&= o(T^{b-1}). \qquad (67)
\end{aligned}$$

We consider another event:

$$\mathbb{P}_{\boldsymbol{\sigma}} \left\{ N_{1,1}(T) < (1 - \delta)\frac{\ln T}{\mathrm{I}(\psi_1, \sigma)}; T_1(T) \geq (1 - c)\pi_1 T \right\}$$

$$\leq \mathbb{P}_{\boldsymbol{\sigma}} \left\{ T_1(T) - N_{1,1}(T) \geq T_1(T) - (1 - \delta)\frac{\ln T}{\mathrm{I}(\psi_1, \sigma)}; \right.$$

$$\left. T_1(T) \geq (1 - c)\pi_1 T \right\}$$

$$\leq \mathbb{P}_{\boldsymbol{\sigma}} \left\{ T_1(T) - N_{1,1}(T) \geq (1 - c)\pi_1 T - (1 - \delta)\frac{\ln T}{\mathrm{I}(\psi_1, \sigma)}; \right.$$

$$T_1(T) \geq (1-c)\pi_1 T\}$$

$$\leq \mathbb{P}_{\boldsymbol{\sigma}}\left\{T_1(T) - N_{1,1}(T) \geq (1-c)\pi_1 T - (1-\delta)\frac{\ln T}{\mathrm{I}(\psi_1,\sigma)}\right\}$$

$$\leq \frac{\mathbb{E}_{\boldsymbol{\sigma}}[T_1(T) - N_{1,1}(T)]}{(1-c)\pi_1 T - O(\ln T)} \qquad \text{(Markov's Inequality)}$$

$$= o(T^{b-1}). \qquad (68)$$

For the event $\mathbb{P}_{\boldsymbol{\sigma}}\left\{N_1(T) < (1-\delta)\frac{\ln T}{\mathrm{I}(\psi_1,\sigma)}\right\}$, we have

$$\mathbb{P}_{\boldsymbol{\sigma}}\left\{N_1(T) < (1-\delta)\frac{\ln T}{\mathrm{I}(\psi_1,\sigma)}\right\}$$

$$\leq \mathbb{P}_{\boldsymbol{\sigma}}\left\{N_{1,1}(T) < (1-\delta)\frac{\ln T}{\mathrm{I}(\psi_1,\sigma)}\right\}$$

$$\leq \mathbb{P}_{\boldsymbol{\sigma}}\left\{T_1(T) \leq (1-c)\pi_1 T\right\}$$

$$+ \mathbb{P}_{\boldsymbol{\sigma}}\left\{N_{1,1}(T) < (1-\delta)\frac{\ln T}{\mathrm{I}(\psi_1,\sigma)}; T_1(T) \geq (1-c)\pi_1 T\right\}$$

$$= o(T^{b-1}). \qquad (69)$$

Note that the allocation rule $\mathcal{L}$ only knows the channel gain realizations of the arms it has played, it does not have the exact distributional knowledge. Let $Y_1, Y_2, \cdots$ denote the successive realizations of the 1-st channel's gains. We define $L_m = \sum_{k=1}^{m} \ln\frac{g(Y_k;\psi_1)}{g(Y_k;\sigma)}$ and an event $\mathcal{E}_T$ as

$$\mathcal{E}_T = \left\{N_1(T) < (1-\delta)\frac{\ln T}{\mathrm{I}(\psi_1,\sigma)}; L_{N_1(T)} \leq (1-b)\ln T\right\}. \qquad (70)$$

From the inequality in (69), we have

$$\mathbb{P}_{\boldsymbol{\sigma}}\left\{\mathcal{E}_T\right\} = o(T^{b-1}). \qquad (71)$$

Note the following relationship

$$\mathbb{P}_{\boldsymbol{\sigma}}\left\{N_1(T) = n_1, \cdots, N_M(T) = n_M; L_{n_1} \leq (1-b)\ln T\right\}$$

$$= \int_{\left\{N_1(T)=n_1, \cdots, N_M(T)=n_M; L_{n_1} \leq (1-b)\ln T\right\}} \prod_{k=1}^{n_1} \frac{g(Y_k;\sigma)}{g(Y_k;\psi_1)} d\mathbb{P}_{\boldsymbol{\psi}}$$

$$= \int_{\left\{N_1(T)=n_1, \cdots, N_M(T)=n_M; L_{n_1} \leq (1-b)\ln T\right\}} e^{-L_{n_1}} d\mathbb{P}_{\boldsymbol{\psi}}$$

$$\geq e^{-(1-b)\ln T}\mathbb{P}_{\boldsymbol{\psi}}\{N_1(T) = n_1, \cdots, N_M(T) = n_M;$$
$$L_{n_1} \leq (1-b)\ln T\}$$

$$= T^{-(1-b)}\mathbb{P}_{\boldsymbol{\psi}}\{N_1(T) = n_1, \cdots, N_M(T) = n_M;$$
$$L_{n_1} \leq (1-b)\ln T\}. \qquad (72)$$

This result rests on the assumption that the allocation rule $\mathcal{L}$ can only depend on the channel gain realizations it has observed by playing and possibly on some internal randomization in the rule. Note that $\mathcal{E}_T$ is a disjoint union of the events of the form $\{N_1(T) = n_1, \cdots, N_M(T) = n_M$ and $L_{n_1} \leq (1-b)\ln T\}$ with $n_1 + \cdots + n_M = T$ and $n_1 < (1-\delta)\frac{\ln T}{\mathrm{I}(\psi_1,\sigma)}$. It now follows from equations (71) and (72) that as $T \to \infty$:

$$\mathbb{P}_{\boldsymbol{\psi}}\{\mathcal{E}_T\} \leq T^{1-b}\mathbb{P}_{\boldsymbol{\sigma}}\{\mathcal{E}_T\} \to 0. \qquad (73)$$

By the strong law of large numbers, $\frac{L_m}{m} \to \mathrm{I}(\psi_1,\sigma) > 0$ and $\max_{k \leq m} \frac{L_k}{m} \to \mathrm{I}(\psi_1,\sigma)$ almost surely under $\mathbb{P}_{\boldsymbol{\psi}}$. Since $1-b >$

$1-\delta$, it follows that as $T \to \infty$:

$$\mathbb{P}_{\boldsymbol{\psi}}\left\{L_k > (1-b)\ln T \text{ for some } k < (1-\delta)\frac{\ln T}{\mathrm{I}(\psi_1,\sigma)}\right\} \to 0. \qquad (74)$$

From equations (73) and (74), we conclude that

$$\lim_{T\to\infty} \mathbb{P}_{\boldsymbol{\psi}}\left\{N_1(T) < \frac{(1-\delta)\ln T}{\mathrm{I}(\psi_1,\sigma)}\right\} = 0.$$

In other words,

$$\lim_{T\to\infty} \mathbb{P}_{\boldsymbol{\psi}}\left\{N_1(T) < \frac{(1-\delta)\ln T}{(1+\delta)\mathrm{I}(\psi_1,\psi_2)}\right\} = 0.$$

This implies that

$$\liminf_{T\to\infty} \mathbb{E}_{\boldsymbol{\psi}}\left[\frac{N_1(T)}{\ln T}\right] \geq \frac{1}{\mathrm{I}(\psi_1,\psi_2)}. \qquad (75)$$

Note that we only considered one optimal arm above. Results like equation (75) hold for all optimal arms. By combining the lower bounds for a fixed non-optimal arm, we get the result in equation (63). ∎

Theorem 9. *Assume that the density and the parameter set satisfy assumptions A1-A4. Let ψ denote the parameter vector whose j-th entry is ψ_j and $\mathcal{O}$ denote the index set of optimal channels. Let $\mathcal{L}$ be any allocation rule that satisfies for every ψ as $T \to \infty$, $\mathfrak{R}_{\mathcal{L}}(T) = o(T^b)$ for every $b > 0$ over an MDP $\mathcal{M}$. Then the regret of $\mathcal{L}$ satisfies*

$$\liminf_{T\to\infty} \frac{\mathfrak{R}_{\mathcal{L}}(T)}{\ln T} \geq \Delta_3 \sum_{i\in\overline{\mathcal{O}}}\left(\max_{j\in\mathcal{O}} \frac{1}{\mathrm{I}(\psi_i,\psi_j)}\right). \qquad (76)$$

Proof: Define a hypothetical allocation $\mathcal{L}'$ based on $\mathcal{L}$ such that whenever $\mathcal{L}$ plays a non-optimal channel for some power level during its execution, $\mathcal{L}'$ plays the optimal channel corresponding to the same power level. It follows $\mathcal{L}$ in rest of the slots. If N' denotes the count variables corresponding to $\mathcal{L}'$, then for $i \in \overline{\mathcal{O}}$ we have $N_i'(T) = 0$ and for $j \in \mathcal{O}$ we have

$$N_j'(T) = N_j(T) + \sum_{i\in\overline{\mathcal{O}}} N_{i,j}(T). \qquad (77)$$

According to equation (41), Δ_3 is the minimum expected gap between the optimal rate of any power level and the rate for any other channel at that power. Using Δ_3, we relate the regrets of $\mathcal{L}$ and $\mathcal{L}'$ as

$$\mathfrak{R}_{\mathcal{L}}(T) \geq \mathfrak{R}_{\mathcal{L}'}(T) + \mathbb{E}\left[\sum_{i\in\overline{\mathcal{O}}} N_i(T)\right]\Delta_3$$

$$\geq \Delta_3 \sum_{i\in\overline{\mathcal{O}}} \mathbb{E}[N_i(T)]. \qquad (78)$$

Using theorem 8, we get the equation (76). ∎

Theorem 9 implies that when the gain distributions characterized by a single parameter for each channel and follow assumptions A1-A4, any algorithm with $\mathfrak{R}(T) = o(T^b)$ for every $b > 0$ must play the non-optimal channels at least $\Omega(\ln T)$ times asymptotically. In the presence of non-optimal channels, an asymptotic regret of $\Omega(\ln T)$ is inevitable for any algorithm. Hence we conclude that our MC-LPSM algorithm is asymptotically order optimal when the system contains non-optimal channels.
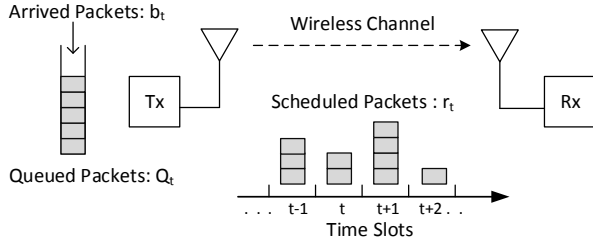
Fig. 2.   Packet scheduling over a wireless channel

VI. COST MINIMIZATION PROBLEMS

We have considered reward maximization problems for describing our online learning framework. This framework can also be applied to average cost minimization problems in packet scheduling with power-delay tradeoff as shown in figure 2. We describe this motivating example and the minor changes required in our algorithms.

Consider a time-slotted communication system where a sender sends data packets to a receiver over a stochastically varying channel with unknown distribution. Such a communication system has been studied previously in [27] assuming the channel to be non-stochastically varying over time. In our setting, the arrival of data packets is also stochastic with a known distribution. The sender can send multiple packets at a higher cost, or can defer some for latter slots while incurring a fixed delay penalty per packet for every time-slot it spends in the sender's queue. Let Q_t denote the number of packets in the queue at time t and $r_t (\leq Q_t)$ be the number of packets transmitted by the sender during the slot. Hence, $Q_t - r_t$ number of packets get delayed. The sender's queue gets updated as

$$Q_{t+1} = \min\{Q_t - r_t + b_t, Q_{\max}\}, \qquad (79)$$

where b_t is the number of new packet arrivals in t-th slot and $Q_{\max}$ is the maximum queue size possible. Since the data-rate is modelled according to equation (1), the power cost incurred during the t-th slot by transmitting r_t packets over the channel becomes $w_p X_t 2^{r_t/B}$, where w_p is a constant known to the sender and X_t is the instantaneous channel gain-to-noise ratio that is assumed to be i.i.d. over time. Assuming w_d as the unit delay penalty, during the slot the sender incurs an effective cost

$$C_t = w_d(Q_t - r_t) + w_p X_t 2^{r_t/B}. \qquad (80)$$

This problem also represents an MDP where the queue size is the state and the number of packets transmitted is the action taken. The goal of this problem is to schedule transmissions r_t sequentially and minimize the expected average cost over time

$$\lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} C_t\right]. \qquad (81)$$

Note that the cost from equation (80) used in this scenario is also a function of the state unlike the problem of energy harvesting communications.

The presented algorithms LPSM and Epoch-LPSM also apply to cost minimization problems with minor changes. If $\rho(\beta, \mathbf{M})$ denotes the average expected cost of the policy β, then $\rho^* = \min_{\beta\in\mathcal{B}} \rho(\beta, \mathbf{M})$. Using this optimal mean cost as the benchmark, we define the cumulative regret of a learning algorithm after T time-slots as

$$\mathfrak{R}(T) := \mathbb{E}\left[\sum_{t=0}^{T-1} C_t\right] - T\rho^*. \qquad (82)$$

In order to minimize the regret for this problem, the LP from (10) needs to be changed from a maximization LP to a minimization LP. With these changes to the algorithms, all the theoretical guarantees still hold with the constants defined accordingly.

VII. NUMERICAL SIMULATIONS

We perform simulations for the power allocation problem with $\mathcal{S} = \{0, 1, 2, 3, 4\}$ and $\mathcal{A} = \{0, 1, 2, 3, 4\}$. Note that each state s_t corresponds to Q_t from equation (2) with $Q_{\max} = 4$ and a_t corresponds to the transmit power q_t from equation (1). The reward function is the rate function from equation (1) and the channel gain is a scaled Bernoulli random variable with $\Pr\{X = 10\} = 0.2$ and $\Pr\{X = 0\} = 0.8$. The valid actions $\mathcal{A}_s$ and the optimal action $\beta^*(s)$ for each state s are shown in table I. We use CVXPY [24] for solving the LPs in our algorithms. For the simulations in figure 3, we use $n_0 = 2$ and $\eta = 10$, and plot the average regret performance over 10^3 independent runs of different algorithms along with their corresponding regret upper bounds. Here, the naive policy never uses the battery, i.e. it uses all the arriving power for the current transmission. Playing such a fixed non-optimal policy causes linearly growing regret over time. Note that the optimal policy also incurs a regret because of the corresponding Markov chain not being at stationarity. We observe that LPSM follows the performance of the optimal policy with the difference in regret stemming from the first few time-slots when the channel statistics are not properly learnt and thus LPSM fails to find the optimal policy. As the time progresses, LPSM finds the optimal policy and its regret follows the regret pattern of the optimal policy. In Epoch-LPSM with $n_0 = 2$ and $\eta = 10$, the agent solves the LP at $t = 1$ and $t = 2$. Its LP solution at $t = 2$ is followed for the first epoch and thus the regret grows linearly till $t = 19$. At $t = 20$, a new LP is solved which often leads to the optimal policy and the regret contribution from latter slots, therefore, follows the regret of the optimal policy. It must be noted that Epoch-LPSM solves only 3 LPs during these slots, while LPSM solves 99 LPs. Epoch-LPSM, therefore, reduces the computational requirements substantially while incurring a slightly higher cumulative regret.

In figure 4, we plot the average regret performance over 10^4 independent runs of Epoch-LPSM for the previous system with different n_0 and η value pairs. As the value of η increases for a fixed n_0, the length of the epochs increases and thus potential non-optimal policies are followed for longer epochs. Larger the value of n_0, better are the policies played in the initial slots where the agent solves the LP at each time. These
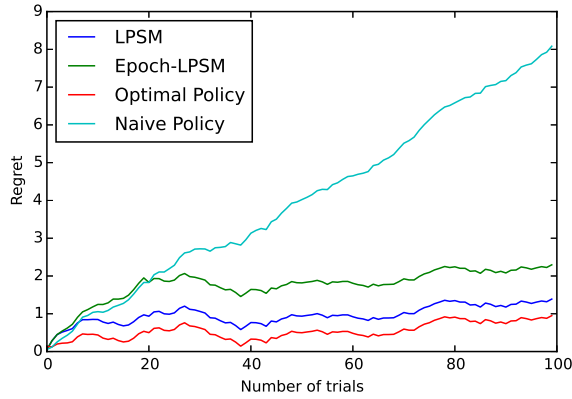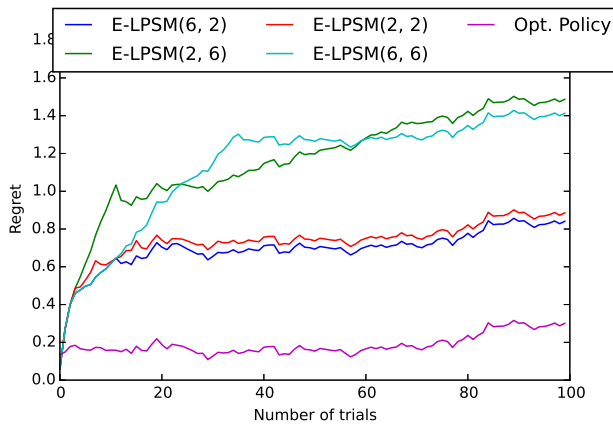
Fig. 3.   Regret performance of LPSM algorithms.

Fig. 4.   Effect of the parameters n_0 and η on the regret of Epoch-LPSM.

TABLE I
ACTIONS FOR EACH STATE

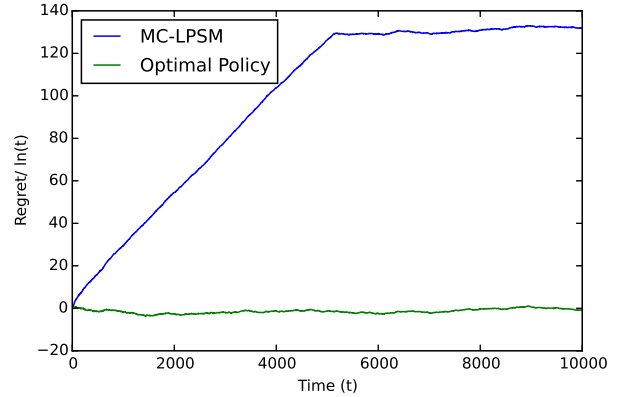| s	$\mathcal{A}_s$	$\beta^*(s)$
0	$\{0\}$	0
1	$\{1\}$	1
2	$\{1, 2\}$	1
3	$\{1, 2, 3\}$	2
4	$\{1, 2, 3, 4\}$	3

Fig. 5.   Regret performance of the MC-LPSM algorithm.

during exploration. As the exploration need gets satisfied over time, the agent solves the LP based on its rate estimates from exploration phases and the regret contribution from the exploitation remains bounded. The regret divided by the logarithm of time saturates, as expected, to a constant value smaller than the asymptotic upper bound of the regret.

VIII. CONCLUSION

We have considered the problem of power allocation over a stochastically varying channel with unknown distribution in an energy harvesting communication system. We have cast this problem as an online learning problem over an MDP. If the transition probabilities and the mean rewards associated with the MDP are known, the optimal policy maximizing the average expected reward over time can be determined by solving an LP specified in the paper. Since the agent is only assumed to know the distribution of the harvested energy, it needs to learn the rewards of the state-action pairs over time and make its decisions based on the learnt behaviour. For this problem, we have proposed two online learning algorithms: LPSM and Epoch-LPSM, which both solve the LP using the sample mean estimates of the rewards instead of the unknown mean rewards. The LPSM algorithm solves the LP at each time-slot using the updated estimates, while the Epoch-LPSM only solves the LP at certain pre-defined time-slots parametrized by n_0 and η and thus, saves a lot of computation at the cost of an increased regret. We have shown that the regrets incurred by both these algorithms are bounded from above by constants. The system designers can, therefore, analyze the regret versus computation tradeoff and tune the parameters n_0 and η based on their performance requirements. Through the numerical simulations, we have shown that the regret of LPSM is very close to that of the optimal policy. We

intuitions are consistent with figure 4, where Epoch-LPSM with $n_0 = 6$ and $\eta = 2$ has the lowest regret and the one with $n_0 = 2$ and $\eta = 6$ has the highest of the lot. We see the regret vs computation tradeoff in action, as the decrease in computation by increasing η or decreasing n_0 leads to larger regrets. We observe that η has more impact on the regret than n_0. Notice that there are changes in the regret trends of Epoch-LPSM at $t = n_0\eta^m$ for small m, because these are the slots where a new LP is solved by MC-LPSM. Once the optimal policy is found by the algorithm, its regret in latter slots follows the trend of the optimal policy.

In figure 5, we plot the regret performance of MC-LPSM for a system with 2 communication channels. The channels are scaled Bernoulli random variables where the gain of the first follows $\Pr\{X = 10\} = 0.5$ and $\Pr\{X = 0\} = 0.5$, while that of the other follows $\Pr\{X = 22\} = 0.4$ and $\Pr\{X = 0\} = 0.6$. The optimal actions selection policy is same as the previous case, while the optimal mapping of transmit-power to the channels is $\phi^*(1) = 2$, $\phi^*(2) = 2$, $\phi^*(3) = 1$ and $\phi^*(4) = 1$. For MC-LPSM, we set $w = 300$ and plot the regret divided by the logarithm of the time index, averaged over 100 realizations, in figure 5. We notice that whenever the exploration slots are densely packed, the regret grows linearly as non-optimal policies are potentially played

have also analyzed the effect of the parameters n_0 and η on the regret the Epoch-LPSM algorithm which approaches the regret of the optimal policy for small η values and large n_0.

For the case of multiple channels, there is an extra layer of decision making to select a channel for transmission in each slot. For this problem, we have extended our approach and proposed the MC-LPSM algorithm. MC-LPSM separates the exploration of different channels to learn their rates from the exploitation where these rate-estimates of different channels for different channels are used to obtain a power selection policy and a channel selection policy in each slot. We have proved a regret upper bound that scales logarithmically in time and linearly in the number of channels. We have also shown that the total computational requirement of MC-LPSM also scales similarly. In order to show the asymptotic order optimality of our MC-LPSM algorithm, we have proved an asymptotic regret lower bound of $\Omega(\ln T)$ for any algorithm under certain conditions. In this paper, we have considered uniform exploration of different channels in MC-LPSM. Analyzing upper confidence bound algorithms, where exploration of different channels gets tuned to their performance, remains a future work.

While we have considered the reward maximization problem in energy harvesting communications for our analysis, we have shown that these algorithms also work for the cost minimization problems in packet scheduling with minor changes.

APPENDIX A
TECHNICAL LEMMAS

Lemma 1 (Hoeffding's Concentration Inequality from [28]). *Let $Y_1, ..., Y_n$ be i.i.d. random variables with mean μ and range $[0, 1]$. Let $S_n = \sum_{t=1}^{n} Y_t$. Then for all $\alpha \geq 0$*

$$\Pr\{S_n \geq n\mu + \alpha\} \leq e^{-2\alpha^2/n}$$
$$\Pr\{S_n \leq n\mu - \alpha\} \leq e^{-2\alpha^2/n}.$$

APPENDIX B
ANALYSIS OF MARKOV CHAIN MIXING

We briefly introduce the tools required for the analysis of Markov chain mixing (see [29], chapter 4 for a detailed discussion). The total variation (TV) distance between two probability distributions ϕ and ϕ' on sample space Ω is defined by

$$\|\phi - \phi'\|_{\text{TV}} = \max_{\mathcal{E} \subset \Omega} |\phi(\mathcal{E}) - \phi'(\mathcal{E})|. \quad (83)$$

Intuitively, it means the TV distance between ϕ and ϕ' is the maximum difference between the probabilities of a single event by the two distributions. The TV distance is related to the L_1 distance as follows

$$\|\phi - \phi'\|_{\text{TV}} = \frac{1}{2} \sum_{\omega \in \Omega} |\phi(\omega) - \phi'(\omega)|. \quad (84)$$

We wish to bound the maximal distance between the stationary distribution π and the distribution over states after t steps of a Markov chain. Let $P^{(t)}$ be the t-step transition matrix

with $P^{(t)}(s, s')$ being the transition probability from state s to s' of the Markov chain in t steps and $\mathcal{P}$ be the collection of all probability distributions on Ω. Also let $P^{(t)}(s, \cdot)$ be the row or distribution corresponding to the initial state of s. Based on these notations, we define a couple of useful t-step distances as follows:

$$d(t) := \max_{s \in \mathcal{S}} \|\pi - P^{(t)}(s, \cdot)\|_{\text{TV}}$$
$$= \sup_{\phi \in \mathcal{P}} \|\pi - \phi P^{(t)}\|_{\text{TV}}, \quad (85)$$
$$\hat{d}(t) := \max_{s, s' \in \mathcal{S}} \|P^{(t)}(s', \cdot) - P^{(t)}(s, \cdot)\|_{\text{TV}}$$
$$= \sup_{\phi', \phi \in \mathcal{P}} \|\phi' P^{(t)} - \phi P^{(t)}\|_{\text{TV}}. \quad (86)$$

For irreducible and aperiodic Markov chains, the distances $d(t)$ and $\hat{d}(t)$ have following special properties:

Lemma 2 ([29], lemma 4.11). *For all $t > 0$, $d(t) \leq \hat{d}(t) \leq 2d(t)$.*

Lemma 3 ([29], lemma 4.12). *The function $\hat{d}$ is sub-multiplicative: $\hat{d}(t_1 + t_2) \leq \hat{d}(t_1)\hat{d}(t_2)$.*

These lemmas lead to following useful corollary:

Corollary 1. *For all $t \geq 0$, $d(t) \leq \hat{d}(1)^t$.*

Consider an MDP with optimal stationary policy β^*. Since the MDP might not start at the stationary distribution π^* corresponding to the optimal policy, even the optimal policy incurs some regret as defined in equation (9). We characterize this regret in the following theorem. Note that this regret bound is independent of the initial distribution over the states.

Theorem 10 (Regret of Optimal Policy). *For an ergodic MDP, the total expected regret of the optimal stationary policy with transition probability matrix P_* is upper bounded by $(1-\gamma)^{-1}\mu_{\max}$, where $\gamma = \max_{s,s' \in \mathcal{S}} \|P_*(s', \cdot) - P_*(s, \cdot)\|_{\text{TV}}$ and $\mu_{\max} = \max_{s \in \mathcal{S}, a \in \mathcal{A}} \mu(s, a)$.*

Proof: Let ϕ_0 be the initial distribution over states and $\phi_t = \phi_0 P_*^{(t)}$ be such distribution at time t represented as a row vectors. Also, let μ^* be a row vector with the entry corresponding to state s being $\mu(s, \beta^*(s))$. We use $d^*(t)$ and $\hat{d}^*(t)$ to denote the t-step distances from equations (85) and (86) for the optimal policy. Ergodicity of the MDP ensures that the Markov chain corresponding to the optimal policy is irreducible and aperiodic, and thus lemmas 2 and 3 hold. The regret of the optimal policy, therefore, gets simplified as:

$$\mathfrak{R}^*(\phi_0, T) = T\rho^* - \sum_{t=0}^{T-1} \phi_t \cdot \mu^*$$
$$= T(\pi^* \cdot \mu^*) - \sum_{t=0}^{T-1} \phi_t \cdot \mu^*$$
$$= \sum_{t=0}^{T-1} (\pi^* - \phi_t) \cdot \mu^*$$
$$\leq \sum_{t=0}^{T-1} (\pi^* - \phi_t)_+ \cdot \mu^*$$

(Negative entries ignored)

18

$$= \sum_{t=0}^{T-1} \sum_{s \in \mathcal{S}} (\pi^*(s) - \phi_t(s))_+ \mu^*(s)$$

$$\leq \mu_{\max} \sum_{t=0}^{T-1} \sum_{s \in \mathcal{S}} (\pi^*(s) - \phi_t(s))_+$$

$$= \mu_{\max} \sum_{t=0}^{T-1} \|\pi^* - \phi_0 P_*^{(t)}\|_{\mathrm{TV}}$$

$$\leq \mu_{\max} \sum_{t=0}^{T-1} d^*(t)$$

$$\leq \mu_{\max} \sum_{t=0}^{T-1} \left(\hat{d}^*(1)\right)^t \qquad \text{(From corollary 1)}$$

$$= \mu_{\max} \sum_{t=0}^{T-1} \gamma^t$$

$$\leq \mu_{\max} \frac{1}{1-\gamma}.$$

$\blacksquare$

References

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

[2] J. Yang and S. Ulukus, "Optimal packet scheduling in an energy harvesting communication system," *Communications, IEEE Transactions on*, vol. 60, no. 1, pp. 220–230, 2012.

[3] K. Tutuncuoglu and A. Yener, "Optimum transmission policies for battery limited energy harvesting nodes," *Wireless Communications, IEEE Transactions on*, vol. 11, no. 3, pp. 1180–1189, 2012.

[4] 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.

[5] 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.

[6] A. Sinha, "Optimal power allocation for a renewable energy source," in *Communications (NCC), 2012 National Conference on*, pp. 1–5, IEEE, 2012.

[7] Z. Wang, A. Tajer, and X. Wang, "Communication of energy harvesting tags," *IEEE Transactions on Communications*, vol. 60, no. 4, pp. 1159–1166, 2012.

[8] C. K. Ho and R. Zhang, "Optimal energy allocation for wireless communications with energy harvesting constraints," *Signal Processing, IEEE Transactions on*, vol. 60, no. 9, pp. 4808–4818, 2012.

[9] S. Vakili, K. Liu, and Q. Zhao, "Deterministic sequencing of exploration and exploitation for multi-armed bandit problems," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 7, no. 5, pp. 759–767, 2013.

[10] 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.

[11] A. Seyedi and B. Sikdar, "Energy efficient transmission strategies for body sensor networks with energy harvesting," *IEEE Transactions on Communications*, vol. 58, no. 7, pp. 2116–2126, 2010.

[12] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[13] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[14] L. Li, W. Chu, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proceedings of the 19th international conference on World wide web*, pp. 661–670, ACM, 2010.

[15] J. Langford and T. Zhang, "The epoch-greedy algorithm for multi-armed bandits with side information," in *Advances in neural information processing systems*, pp. 817–824, 2008.

[16] M. Dudik, D. Hsu, S. Kale, N. Karampatziakis, J. Langford, L. Reyzin, and T. Zhang, "Efficient optimal learning for contextual bandits," in *Conference on Uncertainty in Artificial Intelligence*, 2011.

[17] A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. E. Schapire, "Taming the monster: A fast and simple algorithm for contextual bandits," in *International Conference on Machine Learning*, pp. 1638–1646, 2014.

[18] P. Sakulkar and B. Krishnamachari, "Stochastic contextual bandits with known reward functions." USC ANRG Technical Report, ANRG-2016-02, http://anrg.usc.edu/www/papers/DCB_ANRG_TechReport.pdf.

[19] P. Ortner and R. Auer, "Logarithmic online regret bounds for undiscounted reinforcement learning," in *Proceedings of the 2006 Conference on Advances in Neural Information Processing Systems*, vol. 19, p. 49, 2007.

[20] P. Auer, T. Jaksch, and R. Ortner, "Near-optimal regret bounds for reinforcement learning," in *Advances in neural information processing systems*, pp. 89–96, 2009.

[21] A. Tewari and P. L. Bartlett, "Optimistic linear programming gives logarithmic regret for irreducible mdps," in *Advances in Neural Information Processing Systems*, pp. 1505–1512, 2008.

[22] S. Ross, *Introduction to stochastic dynamic programming*. Academic Press, 1983.

[23] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2005.

[24] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *Journal of Machine Learning Research*, 2016. To appear.

[25] P. Sakulkar and B. Krishnamachari, "Online learning of power allocation policies in energy harvesting communications," in *Signal Processing and Communications (SPCOM), 2016 IEEE International Conference on*, IEEE, 2016.

[26] A. N. Burnetas and M. N. Katehakis, "Optimal adaptive policies for markov decision processes," *Mathematics of Operations Research*, vol. 22, no. 1, pp. 222–255, 1997.

[27] Y. Wu, R. Kannan, and B. Krishnamachari, "Efficient scheduling for energy-delay tradeoff on a time-slotted channel," tech. rep., University of Southern California, Ming Hsieh Department of Electrical Engineering – Systems, 2015.

[28] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American statistical association*, vol. 58, no. 301, pp. 13–30, 1963.

[29] D. A. Levin, Y. Peres, and E. L. Wilmer, *Markov chains and mixing times*. American Mathematical Soc., 2009.

Pranav Sakulkar received his B.Tech. in Electrical Engineering at the Indian Institute of Technology Kanpur in 2011 and his M.Sc. in Communications Engineering at RWTH Aachen University, Germany in 2014. He is a recipient of the Graduate Ph.D. Fellowship and is currently pursuing his Ph.D. in the Department of Electrical Engineering at the University of Southern California's Viterbi School of Engineering. His primary research interest lies in the applications of online learning and bandit problems in wireless networks.

Bhaskar Krishnamachari (Senior Member, IEEE) received his B.E. in Electrical Engineering at The Cooper Union, New York, in 1998, and his M.S. and Ph.D. degrees from Cornell University in 1999 and 2002 respectively. He is a Ming Hsieh Faculty Fellow and Professor in the Department of Electrical Engineering at the University of Southern California's Viterbi School of Engineering. His primary research interest is in the design, theoretical analysis and experimental evaluation of algorithms and protocols for next-generation wireless networks including low power wireless sensor networks. He has co-authored more than 200 publications on this topic, including best paper awards at Mobicom (2010), IPSN (2004, 2010), and MSWiM (2006).