Online Learning of Power Allocation Policies in Energy Harvesting Communications

Pranav Sakulkar and Bhaskar Krishnamachari
Ming Hsieh Department of Electrical Engineering
Viterbi School of Engineering
University of Southern California, Los Angeles, CA, USA
{sakulkar, bkrishna}@usc.edu

***Abstract*—We consider the problem of power allocation over a time-varying channel with an unknown distribution in energy harvesting communication systems. In this problem, the transmitter needs to choose its transmit power based on the amount of stored energy in its battery with the goal of maximizing the average rate obtained over time. We model this problem as a Markov decision process (MDP) with the transmitter as the agent, the battery status as the state, the transmit power as the action and the rate obtained as the reward. The average reward maximization problem over the MDP can be solved by a linear program (LP) that uses the transition probabilities for the state-action pairs and their mean rewards to choose a power allocation policy. Since the rewards associated the state-action pairs are unknown, we propose an online learning algorithm called UCLP that learns these rewards and adapts its policy with time. The UCLP algorithm solves the LP at each time-step to choose its policy using the upper confidence bounds on the rewards. We prove that the reward loss or regret incurred by UCLP is upper bounded by a constant.**

***Index Terms*—Energy harvesting communications, Markov decision process (MDP), online learning, contextual bandits.**

I. Introduction

Communication systems where the transmissions are powered by the harvested energy have rapidly emerged as a viable option for the next-generation wireless networks with prolonged lifetime [1]. The performance of such 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. In [2], power allocation policies over a finite time horizon with known channel gain and harvested energy distributions are studied. In [3], a similar problem is analyzed, but the energy arrivals are assumed to be deterministic and known in advance. The algorithms presented in [4] assume the knowledge of energy arrivals and tries to minimize the overall scheduling time for data packets. In our problem, however, the channel gain distribution is unknown and the harvest energy is assumed be stochastically varying with a known distribution. The transmitter has to decide the 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 an 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.

Our problem can also be seen from the lens of contextual bandits. In the standard contextual bandit problems [5], [6], [7], the contexts are assumed to be drawn from an unknown distribution independently over time. In this paper, we model the context transitions by MDPs. The action the agent takes at time t, therefore, affects not only the instantaneous reward but also the context in slot $t+1$. Thus the agent needs to decide the actions with the global objective in mind, i.e. maximizing the average reward over time. It must be noted that the MDP formulation generalizes the standard contextual bandits [8] for the case where the mapping between the context and random instance to reward is a known monotonic function, 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 [9], [10], [11]. The objective for these problems is to maximize the average undiscounted reward over time. In [9], [10], the agent is unaware of the transition probabilities and the rewards corresponding to the state-action pairs. In [11], the agent knows the rewards, but the transition probabilities are still unknown. In our problem, however, the transition probabilities of the MDP can be inferred from the knowledge of the arrival distribution and the action taken from each state. The goal of our problem is to maximize the average reward by learning the rewards for the state-action pairs over time. One additional feature of our problem is that the function mapping the state-action pair and the channel gain to the rate is known to the agent. The reward information revealed after every action can, therefore, be used to infer the rewards for other state-action pairs.

Our contributions in this paper are as follows:

- We formulate the power allocation problem as an online learning problem over an MDP with the goal of maximizing the average reward over time. We prove that the MDP is ergodic and therefore use its corresponding LP formulation to characterize the optimal policy.
- We propose an online learning algorithm UCLP that learns the rewards for the state-action pairs and adapts its policy over time. We characterize the regret contribution from following a non-optimal policy and from not being at stationarity while following an optimal policy, and prove a constant regret upper bound for UCLP.

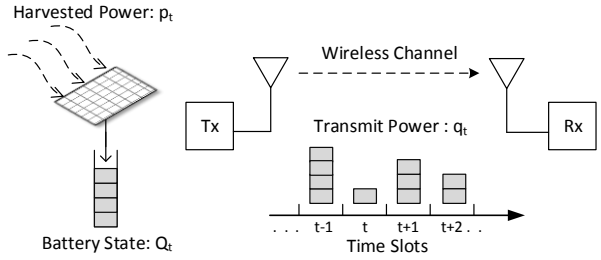

Fig. 1. Power allocation over a wireless channel in energy harvesting communications

This paper is organized as follows. First, we describe the model for the energy harvesting communication system studied in this paper, formulate this problem as an MDP and discuss the structure of the optimal policy in section II. We then propose our online learning algorithm UCLP and prove its regret bounds in section III. Section IV presents the results of numerical simulations for this problem and section V concludes the paper. We also include appendices B and C to discuss and prove some of the technical lemmas at the end of the paper.

II. System Model

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. Let p_t denotes 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

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

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.

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 $a_t \in \mathcal{A}_s$ in state $s_t \in \mathcal{S}$, it receives a random reward $r(s_t, a_t)$. 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

$$r_t(s_t, a_t) = f(s_t, a_t, X_t), \tag{3}$$

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 [12], 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 the following assumptions: 1. when the battery state $Q_t > 0$, the transmit power $q_t > 0$; 2. the distribution of the harvested energy is such that $\Pr\{p_t = p\} > 0$ for all $0 \leq p \leq Q_{\max}$. Under these assumptions, we prove the ergodicity of the MDP as follows.

Proposition 1. *The MDP corresponding to the power allocation application 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,

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

The transition probabilities can, therefore, be expressed as

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

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

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

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. ∎

Since the MDP under consideration is ergodic, we 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

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

For ergodic MDPs, the optimal mean reward ρ^* is independent of the initial state (see [13], section 8.3.3). It is specified as

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

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

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

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. [12], section V.3) shown below.

$$\begin{aligned}
\text{maximize} \quad & \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} \pi(s, a)\mu(s, a) \\
\text{subject to} \quad & \pi(s, a) \geq 0, \ \forall s \in \mathcal{S}, a \in \mathcal{A}_s, \\
& \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} \pi(s, a) = 1, \\
\forall s' \in \mathcal{S}: \quad & \sum_{a \in \mathcal{A}_{s'}} \pi(s', a) = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}_s} \pi(s, a) P(s' \mid s, a),
\end{aligned} \tag{10}$$

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 using standard solvers like CVXPY [14].

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 [13], 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

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

In terms of this notation, the LP from (10) equivalent to $\max_{\beta \in \mathcal{B}} \rho(\beta, \mathbf{M})$. Since the matrix $\mathbf{M}$ is unknown, we develop an online learning framework to learn the optimal policy in the next section.

III. 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 invert them to infer the instantaneous realization of the random variable X. This inference can be used to predict the rewards that would have been obtained 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. Also, we define confidence bounds at time t:

$$u_{t,\lambda}(s, a) = \theta(s, a) + B(s, a)\sqrt{\frac{\lambda \ln t}{t}} \tag{12}$$

$$l_{t,\lambda}(s, a) = \theta(s, a) - B(s, a)\sqrt{\frac{\lambda \ln t}{t}}, \tag{13}$$

which are referred to as UCB and LCB, respectively, and $B(s, a) \geq \max_x f(s, a, x) - \min_x f(s, a, x)$ denotes any upper bound on the maximum possible range of the reward for the state-action pair (s, a). The idea behind our algorithms is to use the UCB values for the maximization problems 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 our online learning algorithm UCLP where the agent solves the LP at each slot. 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.

We use following notations in the analysis of our algorithms: $B_0 := \max_{(s,a)} B(s,a)$, $\Delta_{\min} := \rho^* - \max_{\beta \neq \beta^*} \rho(\beta, \mathbf{M})$. The total number of states and actions are specified as $S := |\mathcal{S}|$, $A := |\mathcal{A}|$, respectively. Also, $\mathbf{U}_{t,\lambda}$ and $\mathbf{L}_{t,\lambda}$ denote the matrices containing the entries $u_{t,\lambda}(s,a)$ and $l_{t,\lambda}(s,a)$ at time t, respectively.

The UCLP algorithm presented in algorithm 1 solves the LP at each time-step and updates it 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 using UCLP. We use this result to bound the total expected regret of UCLP 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, but has no significant impact on the regret bounds.

Algorithm 1 UCLP

Parameters: $\lambda > 1/2$.
Initialization: For all (s,a) pairs, $\theta(s,a) = 0$.
for $n = 0$ **do**
  Given the state s_0 and choose any valid action;
  Update all (s,a) pairs: $\theta(s,a) = f(s,a,x_0)$;
end for
// MAIN LOOP
while 1 **do**
  $n = n + 1$;
  Confidence bounds:

$$u_{n,\lambda}(s,a) = \theta(s,a) + B(s,a)\sqrt{\frac{\lambda \ln n}{n}};$$

  Solve the LP from (10) with $u_{n,\lambda}(s,a)$ instead of unknown $\mu(s,a)$;
  In terms of the LP solution $\pi_{(n)}$, define $\beta_n(s) = \arg\max_{a \in \mathcal{A}_s} \pi_{(n)}(s,a)$, $\forall s \in \mathcal{S}$;
  Given the state s_n, select the action $\beta_n(s_n)$;
  Update for all (s,a) pairs:

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

end while

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

$$n_0 + (1 + A)\, S\sigma_\lambda, \tag{14}$$

where $\sigma_\lambda = \sum_{t=1}^{\infty} t^{-2\lambda}$ *and* n_0 *denotes the minimum value of* $n \in \mathbb{N}$ *for which* $\Delta_{\min} \geq 2B_0\sqrt{\frac{\lambda \ln n}{n}}$.

The proof of theorem 1 is provided in appendix A. UCLP requires $\lambda > 1/2$ in order to have a constant regret upper bound from equation (14), since $\sigma_\lambda < \infty$ for $\lambda > 1/2$. 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 C, we introduce tools to analyze the mixing of Markov chains and characterize this regret contribution in theorem 3. These results are used to upper bound the UCLP regret in the next theorem.

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

$$\Big(n_0 + (1 + A)\, S\sigma_\lambda\Big)\Delta_{\max} + \Big(1 + (1 + A)\, S\sigma_\lambda\Big)\frac{\mu_{\max}}{1 - \gamma}, \tag{15}$$

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 UCLP arises when either non-optimal actions are taken or optimal actions are taken, but 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 C can be applied. According to theorem 3, the regret contribution of any phase is bounded from above by $(1 - \gamma)^{-1}\mu_{\max}$. As proved in theorem 1, for $t \geq n_0$, the expected number of instances of non-optimal policies is upper bounded by $(1 + A)\, S\sigma_\lambda$. Even if none of the these instances appear in successive slots, the expected number of optimal phases is upper bounded by $1 + (1 + A)\, S\sigma_\lambda$. Hence, for $t \geq n_0$, the expected regret contribution from the slots following the optimal policy is upper bounded by

$$\Big(1 + (1 + A)\, S\sigma_\lambda\Big)\frac{\mu_{\max}}{1 - \gamma}. \tag{16}$$

Note that maximum regret possible during one slot is $\Delta_{\max}$. Hence for the first n_0 slots, the regret is bounded by $n_0\Delta_{\max}$. Since there are at most $(1 + A)\, S\sigma_\lambda$ slots with non-optimal policy for $t \geq n_0$ in expectation, their expected regret is upper bounded by $(1 + A)\, S\sigma_\lambda\Delta_{\max}$.

Overall expected regret for the UCLP algorithm is, therefore, bounded from above by equation (15). ∎

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 UCLP, however, we assumed that any occurrence of a non-optimal*

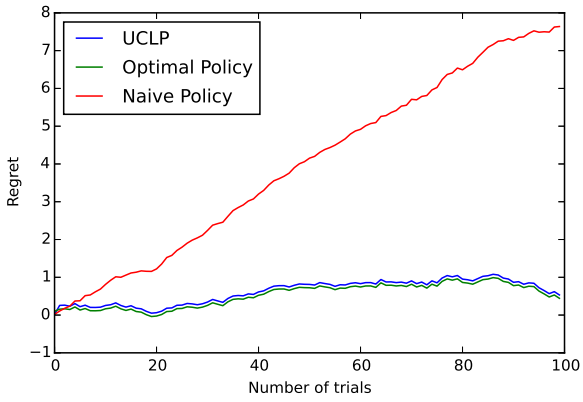

Fig. 2. Regret performance of different algorithms.

policy contributes to the regret. Although this is not necessary, it leads us to a valid upper bound in the proof.

IV. 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 valid actions for each state are shown in table I. 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 $\Pr\{X = 0\} = 0.8$. We use CVXPY [14] for solving the LPs in our algorithm. For the simulations in figure 2, we use $\lambda = 2$ and plot the average regret performance over 10^3 independent runs of different algorithms. Here, the naive policy never uses the battery, i.e. it uses all the arriving power for the current transmission. Note that the optimal policy also incurs a regret because of the corresponding Markov chain not being at stationarity. We observe that UCLP 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 UCLP fails to find the optimal policy. As the time progresses, UCLP finds the optimal policy and the regret follows the regret pattern of the optimal policy.

TABLE I
Valid actions per state

State (s)	Actions ($\mathcal{A}_s$)
0	$\{0\}$
1	$\{1\}$
2	$\{1, 2\}$
3	$\{1, 2, 3\}$
4	$\{1, 2, 3, 4\}$

V. 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 found by solving an LP specified in the paper. Since the agent is only assumed to know the distribution of the harvested energy, she needs to learn the rewards of the state-action pairs over time and make her decisions based on the learnt behaviour. For this problem, we have proposed our online learning algorithm UCLP which solves the LP at each time-slot using the updated upper confidence bounds of the rewards instead of the unknown mean rewards. We have shown that the regret incurred by UCLP is bounded from above by a constant. Through the numerical simulations, we have shown that the regret of UCLP is very close to that of the optimal policy.

The UCLP algorithm presented in this paper solves an LP and therefore requires a lot of computations at each time-step. Reducing the computational needs of UCLP is a potential direction for future works in this area.

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] 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.

[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] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

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

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

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

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

APPENDIX A
PROOF OF THEOREM 1

Let β_t denote the policy obtained by UCLP 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

$$\begin{aligned} N_1 &= 1 + \sum_{t=1}^{\infty} \mathbb{I}\left\{\beta_t \neq \beta^*\right\} \\ &\leq n_0 + \sum_{t=n_0}^{\infty} \mathbb{I}\left\{\beta_t \neq \beta^*\right\} \\ &= n_0 + \sum_{t=n_0}^{\infty} \mathbb{I}\left\{\rho(\beta^*, \mathbf{U}_{t,\lambda}) \leq \rho(\beta_t, \mathbf{U}_{t,\lambda})\right\}. \end{aligned} \tag{17}$$

We observe that $\rho(\beta^*, \mathbf{U}_{t,\lambda}) \leq \rho(\beta_t, \mathbf{U}_{t,\lambda})$ implies that at least one of the following inequalities must be true:

$$\rho(\beta^*, \mathbf{U}_{t,\lambda}) \leq \rho(\beta^*, \mathbf{M}) \tag{18}$$

$$\rho(\beta_t, \mathbf{L}_{t,\lambda}) \geq \rho(\beta_t, \mathbf{M}) \tag{19}$$

$$\rho(\beta^*, \mathbf{M}) < \rho(\beta_t, \mathbf{M}) + \rho(\beta_t, \mathbf{U}_{t,\lambda}) - \rho(\beta_t, \mathbf{L}_{t,\lambda}). \tag{20}$$

Hence we upper bound the probabilities of each of these events. For the first event from condition (18), we get

$$\begin{aligned} &\Pr\left\{\rho(\beta^*, \mathbf{U}_{t,\lambda}) \leq \rho(\beta^*, \mathbf{M})\right\} \\ &= \Pr\left\{\sum_{s\in\mathcal{S}} \pi^*(s, \beta^*(s)) u_{t,\lambda}(s, \beta^*(s)) \right. \\ &\qquad \left. \leq \sum_{s\in\mathcal{S}} \pi^*(s, \beta^*(s)) \mu(s, \beta^*(s))\right\} \\ &\leq \Pr\left\{\text{For at least one state } s \in \mathcal{S} : \right. \\ &\qquad \left. \pi^*(s, \beta^*(s)) u_{t,\lambda}(s, \beta^*(s)) \leq \pi^*(s, \beta^*(s)) \mu(s, \beta^*(s))\right\} \\ &\leq \sum_{s\in\mathcal{S}} \Pr\left\{\pi^*(s, \beta^*(s)) u_{t,\lambda}(s, \beta^*(s)) \right. \\ &\qquad \left. \leq \pi^*(s, \beta^*(s)) \mu(s, \beta^*(s))\right\} \\ &= \sum_{s\in\mathcal{S}} \Pr\left\{u_{t,\lambda}(s, \beta^*(s)) \leq \mu(s, \beta^*(s))\right\} \\ &\overset{(a)}{\leq} \sum_{s\in\mathcal{S}} t^{-2\lambda} \\ &= St^{-2\lambda}, \end{aligned} \tag{21}$$

where (a) holds due to concentration of confidence bounds from lemma 2 (see appendix B).

Similarly for the second event from condition (19), we get

$$\begin{aligned} &\Pr\left\{\rho(\beta_t, \mathbf{L}_{t,\lambda}) \geq \rho(\beta_t, \mathbf{M})\right\} \\ &= \Pr\left\{\sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_s} \pi_{\beta_t}(s,a) l_{t,\lambda}(s,a) \right. \\ &\qquad \left. \geq \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_s} \pi_{\beta_t}(s,a) \mu(s,a)\right\} \\ &\leq \Pr\left\{\text{For at least one state-action pair } (s,a) : \right. \\ &\qquad \left. \pi_{\beta_t}(s,a) l_{t,\lambda}(s,a)) \geq \pi_{\beta_t}(s,a)\mu(s,a)\right\} \\ &\leq \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_s} \Pr\left\{\pi_{\beta_t}(s,a) l_{t,\lambda}(s,a) \geq \pi_{\beta_t}(s,a)\mu(s,a)\right\} \\ &= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_s} \Pr\left\{l_{t,\lambda}(s,a) \geq \mu(s,a)\right\} \\ &\overset{(b)}{\leq} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_s} t^{-2\lambda} \\ &\leq SAt^{-2\lambda}, \end{aligned} \tag{22}$$

where (b) holds due to concentration bounds from lemma 2 (see appendix B).

Now let us analyze the third event from condition (20).

$$\begin{aligned} &\rho(\beta_t, \mathbf{U}_{t,\lambda}) - \rho(\beta_t, \mathbf{L}_{t,\lambda}) \\ &= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_s} \pi_{\beta_t}(s,a) u_{t,\lambda}(s,a) - \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_s} \pi_{\beta_t}(s,a) l_{t,\lambda}(s,a) \\ &= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_s} \pi_{\beta_t}(s,a) \left(u_{t,\lambda}(s,a) - l_{t,\lambda}(s,a)\right) \\ &= \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_s} \pi_{\beta_t}(s,a) \left(2B(s,a)\sqrt{\frac{\lambda \ln t}{t}}\right) \\ &= 2\sqrt{\frac{\lambda \ln t}{t}} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_s} \pi_{\beta_t}(s,a) B(s,a) \\ &\leq 2\sqrt{\frac{\lambda \ln t}{t}} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}_s} \pi_{\beta_t}(s,a) B_0 \\ &\leq 2B_0\sqrt{\frac{\lambda \ln t}{t}}. \end{aligned} \tag{23}$$

Since $\Delta_{\min} \leq \rho(\beta^*, \mathbf{M}) - \rho(\beta_t, \mathbf{M})$, for all $t \geq n_0$ we get

$$\begin{aligned} &\rho(\beta^*, \mathbf{M}) - \rho(\beta_t, \mathbf{M}) - (\rho(\beta_t, \mathbf{U}_{t,\lambda}) - \rho(\beta_t, \mathbf{L}_{t,\lambda})) \\ &\qquad \geq \Delta_{\min} - 2B_0\sqrt{\frac{\lambda \ln t}{t}} \\ &\qquad \geq \Delta_{\min} - 2B_0\sqrt{\frac{\lambda \ln n_0}{n_0}} \\ &\qquad \geq 0. \end{aligned} \tag{24}$$

This implies that condition (20) is always false for $t \geq n_0$.

The expected number of incorrect policies from equation (17), therefore, can be expressed as

$$\begin{aligned} \mathbb{E}[N_1] &\leq n_0 + \sum_{t=n_0}^{\infty} \Pr\left\{\rho(\beta^*, \mathbf{U}_{t,\lambda}) \leq \rho(\beta_t, \mathbf{U}_{t,\lambda})\right\} \\ &\leq n_0 + \sum_{t=n_0}^{\infty} \left(\Pr\left\{\rho(\beta^*, \mathbf{U}_{t,\lambda}) \leq \rho(\beta^*, \mathbf{M})\right\} \right. \\ &\qquad \left. + \Pr\left\{\rho(\beta_t, \mathbf{L}_{t,\lambda}) \geq \rho(\beta_t, \mathbf{M})\right\}\right) \\ &\leq n_0 + \sum_{t=n_0}^{n} \left(St^{-2\lambda} + SAt^{-2\lambda}\right) \\ &\leq n_0 + (1+A)\, S \sum_{t=n_0}^{\infty} t^{-2\lambda} \end{aligned}$$

$$\leq n_0 + (1+A)\, S\sigma_\lambda, \tag{25}$$

where $\sigma_\lambda < \infty$ as $\lambda > 1/2$. ■

APPENDIX B
TECHNICAL LEMMAS & PROOFS

Lemma 1 (Hoeffding's Concentration Inequality from [15])**.** *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 $\alpha \geq 0$*

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

Lemma 2 (Concentration of Confidence Bounds)**.** *At any time t, for any valid state-action pair (s,a), following inequalities hold:*

1) $\Pr\{u_{t,\lambda}(s,a) \leq \mu(s,a)\} \leq t^{-2\lambda}$,
2) $\Pr\{l_{t,\lambda}(s,a) \geq \mu(s,a)\} \leq t^{-2\lambda}$.

Proof: For the first inequality,

$$\begin{aligned}
&\Pr\{u_{t,\lambda}(s,a) \leq \mu(s,a)\} \\
&\leq \Pr\left\{\hat{r}_t(s,a) + B(s,a)\sqrt{\frac{\lambda \ln t}{t}} \leq \mu(s,a)\right\} \\
&= \Pr\left\{\frac{\hat{r}_t(s,a)}{B(s,a)} t \leq \frac{\mu(s,a)}{B(s,a)} t - \sqrt{\lambda t \ln t}\right\} \\
&\overset{(a)}{\leq} \mathrm{e}^{-2(\lambda t \ln t)/t} \\
&= t^{-2\lambda},
\end{aligned} \tag{26}$$

where (a) is obtained using the left-sided Hoeffding's inequality (see lemma 1) with $\alpha = \sqrt{\lambda t \ln t}$.

Similarly using the right-sided version of the concentration inequality, we get the second inequality. ■

APPENDIX C
ANALYSIS OF MARKOV CHAIN MIXING

We briefly introduce the tools required for the analysis of Markov chain mixing (see [16], chapter 4 for a detailed discussion). The total variation (TV) distance between two probability distributions ϕ and ψ on sample space Ω is defined by

$$\|\phi - \psi\|_{\mathrm{TV}} = \max_{\mathcal{E} \subset \Omega} |\phi(\mathcal{E}) - \psi(\mathcal{E})|. \tag{27}$$

Intuitively, it means the TV distance between ϕ and ψ 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

$$\|\phi - \psi\|_{\mathrm{TV}} = \frac{1}{2} \sum_{\omega \in \Omega} |\phi(\omega) - \psi(\omega)|. \tag{28}$$

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:

$$\begin{aligned}
d(t) &:= \max_{s \in \mathcal{S}} \|\pi - P^{(t)}(s,\cdot)\|_{\mathrm{TV}} \\
&= \sup_{\phi \in \mathcal{P}} \|\pi - \phi P^{(t)}\|_{\mathrm{TV}},
\end{aligned} \tag{29}$$

$$\begin{aligned}
\hat{d}(t) &:= \max_{s,s' \in \mathcal{S}} \|P^{(t)}(s',\cdot) - P^{(t)}(s,\cdot)\|_{\mathrm{TV}} \\
&= \sup_{\psi,\phi \in \mathcal{P}} \|\psi P^{(t)} - \phi P^{(t)}\|_{\mathrm{TV}}.
\end{aligned} \tag{30}$$

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

Lemma 3 ([16], lemma 4.11)**.** *For all $t > 0$, $d(t) \leq \hat{d}(t) \leq 2d(t)$.*

Lemma 4 ([16], 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.

Theorem 3 (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 $\gamma = \max_{s,s' \in \mathcal{S}} \|P_*(s',\cdot) - P_*(s,\cdot)\|_{\mathrm{TV}}$ 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 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 (29) and (30) 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 3 and 4 hold. The regret of the optimal policy, therefore, gets simplified as:

$$\begin{aligned}
&\mathfrak{R}^*(\phi_0, T) \\
&= T\rho^* - \sum_{t=0}^{T-1} \phi_t \cdot \mu^* \\
&= T(\pi^* \cdot \mu^*) - \sum_{t=0}^{T-1} \phi_t \cdot \mu^* \\
&= \sum_{t=0}^{T-1} (\pi^* - \phi_t) \cdot \mu^* \\
&\leq \sum_{t=0}^{T-1} (\pi^* - \phi_t)_+ \cdot \mu^* \qquad \text{(Negative entries ignored)}
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{t=0}^{T-1} \sum_{s \in \mathcal{S}} (\pi^*(s) - \phi_t(s))_+ \mu^*(s) \\
&\leq \mu_{\max} \sum_{t=0}^{T-1} \sum_{s \in \mathcal{S}} (\pi^*(s) - \phi_t(s))_+ \\
&= \mu_{\max} \sum_{t=0}^{T-1} \|\pi^* - \phi_0 P_*^{(t)}\|_{\mathrm{TV}} \\
&\leq \mu_{\max} \sum_{t=0}^{T-1} d^*(t) \\
&\leq \mu_{\max} \sum_{t=0}^{T-1} \left(\hat{d}^*(1)\right)^t \qquad \text{(From corollary 1)} \\
&= \mu_{\max} \sum_{t=0}^{T-1} \gamma^t \\
&\leq \mu_{\max} \frac{1}{1-\gamma}.
\end{aligned}
$$

Note that this regret bound is independent of the initial distribution over the states. ■