EFFICIENT ALGORITHMS FOR LINEAR POLYHEDRAL BANDITS

Manjesh K. Hanawal *Amir Leshem*[⋆] *Venkatesh Saligrama*[†]

IEOR Group, IIT-Bombay, Mumbai, India 400076
[⋆] Dept. of EE, Bar-Ilan University, Ramat-Gan, Israel 52900
[†] Dept. of ECE, Boston University, Boston, USA 02215

ABSTRACT

We study stochastic linear optimization problem with bandit feedback. The set of arms take values in an N-dimensional space and belongs to a bounded polyhedron described by finitely many linear inequalities. We present an algorithm that has $O(N \log^{1+\epsilon}(T))$ expected regret for any $\epsilon > 0$ in T rounds. The algorithm alternates between exploration and exploitation phases where it plays a deterministic set of arms in the exploration phases and a greedily selected arm in the exploitation phases. The regret bound of SEE compares well to the lower bounds of $\Omega(N \log T)$ that can be derived by a direct adaptation of Lai-Robbin's lower bound proof [1]. Our key insight is that for a polyhedron the optimal arm is robust to small perturbations in the reward function. Consequently, a greedily selected arm is guaranteed to be optimal when the estimation error falls below a suitable threshold. Our solution resolves a question posed by [2] that left open the possibility of efficient algorithms with logarithmic regret bounds. The simplicity of our approach allows us to derive probability one bounds on the regret, in contrast to the weak convergence results of other papers. This ensures that with probability one only finitely many errors occur in the exploitation phase. Numerical investigations show that while theoretical results are asymptotic the performance of our algorithms compares favorably to state-of-the-art algorithms in finite time as well.

1. INTRODUCTION

Stochastic bandits are sequential decision making problems where a learner plays an action in each round and observes the corresponding reward. The goal of the learner is to collect as much reward as possible or, alternatively minimize regret over a period of T rounds. *Stochastic linear bandits* are a class of *structured bandit problems* where the rewards from different actions are correlated. In particular, the expected reward of each action or arm is expressed as an inner product of a feature vector associated with the action and an unknown parameter which is identical for all the arms. With this structure, one can infer reward of arms that are not yet played from the observed rewards of other arms. This allows for considering cases where playing each arm is infeasible as there number could be large or unbounded.

Stochastic linear bandits have found rich applications in many fields including web advertisements [3], recommendation systems [4], packet routing, revenue management, etc. In many applications the set of actions are often defined by a finite set of constraints. For example, in packet routing, the amount of traffic to be routed on a link is constrained by its capacity. In web-advertisements problems, the budget constraints determine the set of available advertisements. It follows that the set of arms in these applications is a polyhedron.

The stochastic bandit setting has been extensively used in the study channel allocation in cognitive radio networks. In [5] [6], the authors study cognitive radio networks where a secondary user accesses the primary channels that are modeled as arms of a bandit problems. The authors in [7] [8] extend the setting to the scenario where there are multiple secondary users. Adversarial bandits (non-stochastic) [9] bandit setting is applied in the study of conginitve radio networks where the secondary users can sense some channel for activity while they transmit on some channels [10]. Linear stochastic stochastic bandits have also been applied in the study cognitive networks where one can represent the correlation across the channels through a graph [11]. In [12], the authors consider a combinatorial setting where a user can select multiple channels simultaneously and gets linear sum of weighted rewards from each channel.

Bandit algorithms are evaluated by comparing their cumulative reward against the optimal achievable cumulative reward and the difference is referred to as regret. Typically, the machine learning literature distinguishes two types of characterization of the regret performance: The minimax bounds, where the regret performance is evaluated over the worst reward (probability) distributions, and the problem dependent bounds, were the regret performance is evaluated under a fixed (unknown) reward distribution. The focus of this paper is one the later type of performance characterization. The linear bandit problem [2, 13] deals with the case where the rewards of the different arms are linear functions of an unknown parameter vector. For linear bandits, minimax regret lower bounds are well studied and stated in terms of dimension of the set of arms rather than its size. The most commonly studied problem is that of linear bandits over compact sets. For the case where the number of arms is infinite or form a bounded subset of a N-dimensional space, a lower bound of $\Omega(N\sqrt{T})$ is established in [2, 13] where T is the number of rounds. The authors also develop algorithms that have matching upper bounds on the cumulative regret. Several variants and special cases of stochastic linear bandits are available depending on what forms the set of arms. The classical stochastic multi-armed bandits [14], [1] is a special case of linear bandits where the set of actions available in each round is the standard orthonormal basis. For other variants we refer to the technical report [15].

The classical stochastic multi-armed bandits introduced by Robbins [14] and later studied by Lai and Robbins [1] dealt with discrete set of N bandits and established an asymptotic problem dependent lower bound that is logarithmic in T. Linear bandits are an important generalization of multi-armed bandits when the arms are correlated, so that information from one arm provides information on other arms. In this paper we focus on linear bandits where the arms take values in an N-dimensional space and belong to a bounded polyhedron described by finitely many linear inequalities. We derive an asymptotic lower bound of $\Omega(N \log T)$ for this problem and present an algorithm that is (almost) asymptotically optimal. Our so-

lution resolves a question posed by [2] that left open the possibility of efficient algorithms with asymptotic logarithmic regret bounds.

Our algorithm SEE (Sequential Exploration and Exploitation) alternates between exploration and exploitation phases, where a set of arms on the boundary of the polyhedron are played in exploration phases and a greedily selected arm is played super-exponentially many times in the exploitation phases. The simplicity of our approach allows us to derive probability one bounds on the regret, in contrast to the weak convergence results of other papers. This ensures that with probability one only finitely many errors occur in the exploitation phase. The regret of upper confidence bound (UCB) based algorithms concentrates only at a polynomial rate [16]. Thus, our algorithms are more suitable for risk-averse decision making. Numerical experiments show that its regret performance compares well against state-of-the-art linear bandit algorithms even for reasonably small rounds while being significantly better asymptotically.

Related Work: Our regret bounds are related to those described in [13], who present an algorithm ($ConfidenceBall_2$) with regret bounds that scale as $O((N^2/\Delta)\log^3 T)$ with hight probability, where Δ is the reward gap defined over extremal points. This bound is improved to $\mathcal{O}((\log^2 T + N\log T + N^2\log\log T)/\Delta)$ in [17]. These algorithms belong to the class of so called OFU (Optimism in the Face of Uncertainty) algorithms. Since OFU algorithms play only extremal points (arms), one may think that $\log T$ regret bounds can be attained for linear bandits by treating them as K-armed bandits, were K denotes the number of extremal points of the set of actions. This possibility arises from the classical results on the K-armed bandit problem due to Lai and Robbins [1] who provided a complete characterization of expected regret by establishing lower bound of $\Omega(K\log T)$ and then providing an asymptotically (optimal) algorithm with a matching upper bound. But, as noted in [2][Sec 4.1, Example 4.5], the number of extremal points can be exponential in N, and this renders such adaptation of multi-armed bandits algorithm inefficient. In the same paper, the authors pose it as an open problem to develop efficient algorithms for linear bandits over polyhedral set of arms that have logarithmic regret. They also remark that since convex hull of a polyhedron is not strongly convex, regret guarantees of their PEGE (Phased Exploration Greedy Exploitation) algorithm does not hold.

Our work is close to FEL (Forced Exploration for Linear bandits) algorithm developed in [18]. FEL separates the exploration and exploitation phases by comparing the current round number against a predetermined sequence. FEL plays randomly selected arms in the exploration intervals and greedily selected arms in the exploitation intervals. However, our policy differs from FEL as follows– 1) we always play fixed set of arms (deterministic) in the exploration phases. 2) noise is assumed to be bounded in [18], whereas we consider more general sub-Gaussian noise model 3) unlike FEL, our policy does not require computationally costly matrix inversions. FEL provides expected regret guarantee of only $\mathcal{O}(c\log^2 T)$ whereas our policy SEE has almost optimal $\mathcal{O}(N\log^{1+\epsilon} T)$ regret guarantee for any $\epsilon > 0$.

The paper is organized as follows: In Section 2, we describe the problem and setup notations. In Section 3, we derive a lower bound on expected regret and describe our main algorithm SEE. Finally, we numerically compare performance of our algorithm against sate-of-the-art in 5.

2. PROBLEM FORMULATION

We consider a stochastic linear optimization problem with bandit feedback over a set of arms defined by a polyhedron. Let $\mathcal{C} \subset \mathcal{R}^N$ denote a bounded polyhedral given by

$$\mathcal{C} = \left\{\mathbf{x} \in \mathcal{R}^N : \mathbf{A}\mathbf{x} \leq \mathbf{b}\right\} \tag{1}$$

where $\mathbf{A} \in \mathcal{R}^{M\times N}, \mathbf{b} \in \mathcal{R}^M$. At each round t, selecting an arm $x_t \in \mathcal{C}$ results in reward $r_t(\mathbf{x}_t)$. We investigate the case where the expected reward for each arm is a linear function regardless of the history. I.e., for any history $\mathcal{H}_t$, there is a parameter $\boldsymbol{\theta} \in [-1,1]^N$, fixed but unknown, such that

$$\mathbb{E}[r_t(\mathbf{x})|\mathcal{H}_t] = \boldsymbol{\theta}'\mathbf{x} \quad \text{for all } \ t \text{ and } \mathbf{x} \in \mathcal{C}.$$

Under these setting the noise sequence $\{\nu_t\}_{t=1}^{\infty}$, where $\nu_t = r_t(\mathbf{x}) - \mathbf{x}'\boldsymbol{\theta}$ forms a martingale difference sequence. Let $\mathcal{F}_t = \sigma\{\nu_1, \nu_2, \cdots, \nu_t, \mathbf{x}_1, \cdots, \mathbf{x}_{t+1}\}$ denote the σ-algebra generated by noise events and arms selections till time t. Then ν_t is $\mathcal{F}_t$-measurable and we assume that it satisfies

$$\text{for all } b \in \mathcal{R}^1 \quad \mathbb{E}[e^{b\nu_t}|\mathcal{F}_{t-1}] \leq \exp\{b^2R^2/2\}, \tag{2}$$

i.e., noise is conditionally R- sub-Gaussian which automatically implies $\mathbb{E}[\nu_t|\mathcal{F}_t] = 0$ and $\mathbf{Var}(\nu_t) \leq R^2$. We can think of R^2 as the conditional variance of noise. An example of R-sub-Gaussian noise is $\mathcal{N}(0, R^2)$, or any bounded distribution over an interval of length $2R$ and zero mean. In our work, R is fixed but unknown.

A policy $\phi := (\phi_1, \phi_2, \cdots)$ is a sequence of functions $\phi_t : \mathcal{H}_{t-1} \to \mathcal{C}$ such that an arm is selected in round t based on the history $\mathcal{H}_{t-1}$. Define expected (pseudo) regret of policy ϕ over T-rounds as:

$$R_T(\phi) = T\boldsymbol{\theta}'\mathbf{x}^* - E\left[\sum_{t=1}^{T}\boldsymbol{\theta}'\phi(t)\right] \tag{3}$$

where $\mathbf{x}^* = \arg\max_{\mathbf{x}\in\mathcal{C}}\boldsymbol{\theta}'\mathbf{x}$ denotes the optimal arm in $\mathcal{C}$, which exists and is an extremal point[1] of the polyhedron $\mathcal{C}$ [19]. The expectation is over the random realization of the arm selections induced by the noise process. The goal is to learn a policy that keeps the regret as small as possible. We will be also interested in regret of the policy defined as

$$\overline{R}_T(\phi) = T\boldsymbol{\theta}'\mathbf{x}^* - \sum_{t=1}^{T}\boldsymbol{\theta}'\phi(t). \tag{4}$$

For the above setting, we can use $ConfidenceBall_2$ [13] or $UncertainityEllipsoid$ [2] and achieve optimal regret of order $N\sqrt{T}$. For linear bandits over a set with finite number of extremal points, one can also achieve regret that scales more gracefully, growing logarithmically in time T, using algorithms for the standard multi-armed bandits. Indeed, from fundamentals of linear programming

$$\arg\max_{\mathbf{x}\in\mathcal{C}}\boldsymbol{\theta}'\mathbf{x} = \arg\max_{\mathbf{x}\in\mathcal{E}(\mathcal{C})}\boldsymbol{\theta}'\mathbf{x},$$

where $\mathcal{E} := \mathcal{E}(\mathcal{C})$ denotes the set of extremal points of $\mathcal{C}$. Since the set of extremal points is finite for a polyhedron, we can use the standard Lai and Robbin's algorithm [1] or UCB1 in [20] treating each extremal point as an arm and obtain regret bound (problem dependent) of order $\frac{|\mathcal{E}|}{\Delta}\log T$, where $\Delta := \boldsymbol{\theta}'\mathbf{x}^* - \max_{\mathcal{E}\setminus\mathbf{x}^*}\boldsymbol{\theta}'\mathbf{x}$ denotes the gap between the best and the next best extremal point. However, the leading term in these bounds can be exponential in N, rendering these algorithm ineffective. For example, the number of extremal points of $\mathcal{C}$ can be of the order $\binom{M+N}{M} = \mathcal{O}((2N)^M)$. Nevertheless, in analogy with the problem independent regret bounds in linear

[1]Extremal point of a set is a point that is not a proper convex combination of points in the set.

bandits, one wishes to derive problem dependent logarithmic regret where the dependence on set of arms is only linear in its dimension. Hence we seek an algorithm with regret of order $N \log T$.

3. MAIN RESULTS

In the following, we first derive a lower bound on the expected regret and develop an algorithm that is (almost) asymptotically optimal.

3.1. Lower Bound

We derive the lower bound for linear bandits on polyhedral sets following the steps in [15]after identifying an instance of polyhedron and mapping the setting to the N-armed bandit problem. The detailed proof is given in [15]. Without loss of generality, we restrict our attention to uniformly good policies as defined in [1]. We say that a policy ϕ is uniformly optimal if for all $\boldsymbol{\theta} \in \Theta$, $R(T,\phi) = o(T^{\alpha})$ for all $\alpha > 0$.

Theorem 1 *Let ϕ any uniformly good policy on a bounded polyhedron with positive measure. For any $\boldsymbol{\theta} \in [0,1]^N$, let $\mathbb{E}[\eta(\theta_k)] = \theta_k$ for all k and $\theta^* = \arg\max_n \theta_n$. Then,*

$$\liminf_{T\to\infty} \frac{R_T(\phi)}{\log T} \geq \frac{(N-1)\Delta}{\max\limits_{k:\theta_k<\theta^*} KL(\theta^*,\theta_k)} \tag{5}$$

where $KL(\theta^*,\theta_k)$ denotes the Kullback-Leibler divergence between the distributions parametrized by θ^* and θ_k.

3.2. Algorithm

The basic idea underlying our proposed technique is based on the following observations for linear optimization over a polyhedron. 1) The set of extremal points of polyhedron is finite and hence $\Delta > 0$. 2) When $\hat{\boldsymbol{\theta}}$ is sufficiently close to $\boldsymbol{\theta}$, then over the set $\mathcal{C}$ both $\arg\max \boldsymbol{\theta}'\mathbf{x}$ and $\arg\max \hat{\boldsymbol{\theta}}'\mathbf{x}$ give the same value. We exploit these observations and propose a two stage technique, where we first estimate $\boldsymbol{\theta}$ based on a block of samples and then exploit it for much longer block. This is repeated with increasing block lengths so that at each point the regret is logarithmic.

For ease of exposition, we consider a polyhedron $\mathcal{C}$ which contains the origin as an interior point. The method can be extended to general case by using an interior point of the polyhedron as proxy for the origin. The details are provided in the technical report [15].

Let $\mathbf{e}_n$ denote nth standard unit vector of dimension N. For all $1 \leq n \leq N$, let $\overline{z}_n = \max\{z \geq 0, z\mathbf{e}_n \in \mathcal{C}\}$. The subset of arms $\mathcal{B} := \{\overline{z}_n \mathbf{e}_n : n = 1, 2\cdots, N\}$ are the vertices of the largest simplex bounded in $\mathcal{C}$. Since $\theta_n = \boldsymbol{\theta}'\mathbf{e}_n$ we can estimate θ_n by repeatedly playing the arm $\overline{z}_n\mathbf{e}_n$. One can also estimate θ_n by playing an interior point $z\mathbf{e}_n \in \mathcal{C}$ for some $z > 0$. But as will see later selecting the maximum possible z improves the probability of estimation error.

In our policy- which we refer as **S**equential-**E**stimation-**E**xploitation (SEE)- we split the time horizon into cycles and each cycle consists of an exploration interval followed by an exploitation interval. We index the cycles by c and denote the exploration and exploitation intervals in cycle c as E_c and R_c, respectively. In the exploration interval E_c, we play each arm in $\mathcal{B}$ repeatedly for $(2c+1)$ times. At the end of E_c, using the rewards observed for each arm in $\mathcal{B}$ in the past c- cycles we compute ordinary least square (OLS) to estimate each component $\theta_n, n = 1, 2, \cdots, N$ separately and obtain the estimate $\hat{\boldsymbol{\theta}}(c)$. Using $\hat{\boldsymbol{\theta}}(c)$ as a proxy for $\boldsymbol{\theta}$, we compute a greedy arm $\mathbf{x}(c)$ by solving a linear program and play it repeatedly for $2^{c^2/(1+\epsilon)}$ times in the exploitation interval R_c, where $\epsilon > 0$ in an input parameter. We repeat the process for each cycle. A formal description of SEE is given in figure 1. The estimation in line 9 is computed for all $n = 1, 2, \cdots, N$ as follows:

Algorithm 1 SEE

1: **Input:** $\mathcal{C}$: The polyhedron ϵ: Algorithm parameter
2: **Initialization:** Compute the set $\mathcal{B}$
3: **for** $c = 0, 1, 2, \cdots$ **do**
4: **Exploration:**
5: **for** $n = 1 \to N$ **do**
6: **for** $j = 1 \to 2c+1$ **do**
7: Play arm $z_n\mathbf{e}_n \in \mathcal{B}$, observe reward $r_{t_{c,n,j}}$
8: **end for**
9: Compute $\hat{\theta}_n(c)$
10: **end for**
11: $\mathbf{x}(c) \leftarrow \arg\max\limits_{\mathbf{x}\in\mathcal{C}} \mathbf{x}'\hat{\boldsymbol{\theta}}(c)$
12: **Exploitation:**
13: **for** $j = 1 \to \lfloor 2^{c^2/1+\epsilon} \rfloor$ **do**
14: Play arm $\mathbf{x}$(c), observe reward
15: **end for**
16: **end for**

$$\hat{\theta}_n(c) = \frac{1}{(c+1)^2} \sum_{i=0}^{c} \sum_{j=1}^{2i+1} r_{t_{i,n,j}}/z_n, \tag{6}$$

Note that in the exploration intervals, SEE plays a fixed set of arms and no adaption happens, adding positive regret in each cycle. The regret incurred in the exploitation intervals starts reducing as the estimation error gets small, and when it falls below $\Delta/2$ the step (line-11) selects the optimal arm and no regret is incurred in the exploitation intervals (see Lemma 2 in [15]). The probability of selecting the optimal arm decays super-exponentially across the cycles, and hence the probability of incurring positive regret in the exploitation intervals also decays super-exponentially. The SEE provides the following guarantee on the expected regret.

Theorem 2 *Let the noise be R-sub-Gaussian and without loss of generality[2] assume $\boldsymbol{\theta} \in [-1,1]^N$. Then, the expected regret of SEE, with parameter $\epsilon > 0$ is bounded as follows:*

$$R_T(SEE) \leq 2R_m N \log^{1+\epsilon} T + 4R_m N\gamma_1, \tag{7}$$

where R_m denotes the maximum reward. γ_1 is a constant that depends on noise parameter R and the sub-optimality gap Δ.

The ϵ parameter determines the length of the exploitation intervals, and larger ϵ implies that SEE spends less time in exploitation and more time in exploration. Increasing ϵ will make SEE spend more time in explorations resulting in improved estimations and reduces the probability of playing sub-optimal arm in the exploitation intervals. Hence parameter ϵ determines how fast the regret concentrates, and larger its value more 'risk-averse' is the algorithm.

[2]For general $\boldsymbol{\theta}$, we replace it by $\frac{\boldsymbol{\theta}}{\|\boldsymbol{\theta}\|_\infty}$ and the same method works. Only R_m is scaled by a constant factor.

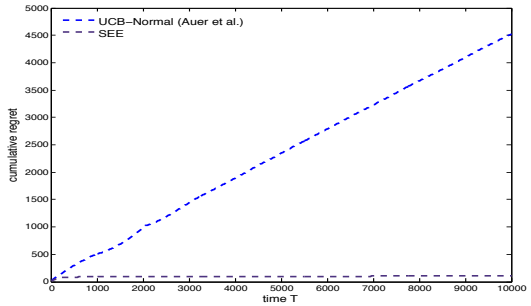

Fig. 1: Regret comparison against multi-armed bandits, arms are corners of 10-dim. hypercube.

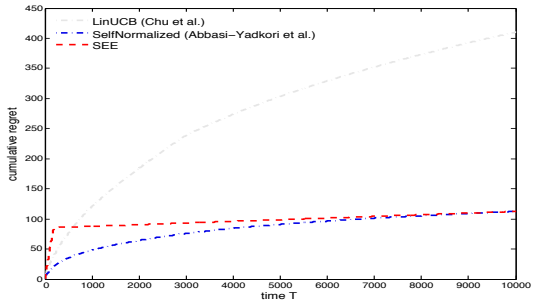

Fig. 2: Regret comparison against linear bandit algorithms on 10-dim. hypercube.

3.3. Regret of SEE.

We analyze the regret in the Exploration and Exploitation phases separately as follows.

Exploration regret: At the end of cycle c, each arm in $\mathcal{B}$ is played $\sum_{i=1}^{c}(2i+1) = c^2$ times. The total expected regret from the exploration intervals after c cycles is at most Nc^2R_m.

Exploitation regret: Total expected regret from the exploration intervals after c cycle is

$$\begin{aligned} 4NR_m \sum_{i=1}^{c} 2^{i^{2/(1+\epsilon)}} 2^{-i^2 a\Delta^2} &= 4NR_m \sum_{i=1}^{c} 2^{i^{2/(1+\epsilon)} - i^2 a\Delta^2} \\ &\leq 4NR_m\gamma_2 \end{aligned}$$

where $\gamma_2 := \sum_{i=1}^{\infty} 2^{i(i^{(1-\epsilon)/(1+\epsilon)} - c_1 i\Delta^2/4)}$ is a convergent series. After c cycles, the total number of plays is $T = \sum_{i=1}^{c} e^{i^{\frac{2}{1+\epsilon}}} + Nc^2 \geq e^{c^{\frac{2}{1+\epsilon}}}$ and we get $c^2 \leq \log^{1+\epsilon} T$. Finally, expected regret form T-rounds is bounded as

$$R_T(SEE) \leq 2R_m N \log^{1+\epsilon} T + 4NR_m\gamma_2 = \mathcal{O}(N \log^{1+\epsilon} T).$$

4. PROBABILITY 1 REGRET BOUNDS.

Recall the definiton of expected regret and regret in (3) and (4). In this section we show that with probability 1, the regret of our algorithms are within a constant factor from the their expected regret.

Theorem 3 *With probability* 1, $\overline{R}_T(SEE)$ *is* $\mathcal{O}(N \log^{1+\epsilon} T)$.

Proof: Let $\mathbb{C}_n$ denote an event that we select sub-optimal arm in the nth cycle. From Lemma 2 in [15], this event is bounded as $\Pr\{\mathbb{C}_n\} \leq N \exp\{-\mathcal{O}(n^2)\}$. Hence $\sum_{n=1}^{\infty} \Pr\{\mathbb{C}_n\} < \infty$. Now, from application of Borel-Cantelli lemma, we get $\Pr\{\limsup_{n\to\infty} \mathbb{C}_n\} = 0$, which implies that almost surely SEE plays optimal arm in all but finitely many cycles. Hence the exploitation intervals contribute only a bounded regret. Since the regret due to exploration intervals is deterministic, the regret of SEE are within a constant factor from their expected regret with probability 1, i.e., $\Pr\{\exists\ C_1$ such that $\overline{R}_T(SEE) \leq R_T(SEE) + C_1\}$. This completes the claim.

We note that the regret bounds proved in [13] hold with high confidence, where as ours hold with probability 1 and hence provides a stronger performance guarantee. This result is not only a mathematical detail. It actually ensures that with probability 1, only finitely many times we use the wrong arm in the exploitation phase.

5. EXPERIMENTS

In this section we investigate numerical performance of our algorithms against the known algorithms. We run the algorithms on a hypercube with dimension $N = 10$. We generated $\boldsymbol{\theta} \in [0,1]^N$ randomly and noise is zero mean Gaussian random variable with variance 1 in each round. The experiments are averaged over 10 runs. In Fig. 1 we compare SEE ($\epsilon = 0.3$) against UCB-Normal [21], where we treated each extremal point as an arm of an 2^N-armed bandit problem. As expected, our algorithms perform much better. UCB-Normal need to sample each of the 2^N atleast once before it could start learning the right arm. Whereas, our algorithm starts playing the right arm after a few cycles of exploration intervals. In Fig. 2, we compare our algorithms against the linear bandits algorithm LinUCB and self-normalization based algorithm in [22], which is labeled SelfNormalized in the figure. For these we set confidence parameter to 0.001. We see that SEE beats LinUCB by a huge margin, but its performance comes close to that of SelfNormalized algorithm. Note that SelfNormalzed algorithm requires knowledge of noise sub-Gaussianity parameter R. Whereas, our algorithms are agnostic to this parameter. Also, note that in the early rounds the performance of SEE is close to SelfNormalized algorithm. However, for large T its peformance is improves compared to SelfNormalized as can be noticed at the edge of the figure. In all the numerical plots, we initialized the algorithm to run from cycle number 5.

6. CONCLUSION

We studied stochastic linear optimization over polyhedral set of arms with bandit feedback. We provided asymptotic lower bound for any policy and developed an algorithm that is near optimal. The regret of the algorithm grows (near) logarithmically in T and its growth rate is linear in the dimension of the polyhedron. We showed that the regret upper bounds hold almost surely. The regret growth rate of our algorithms is $\log^{1+\epsilon} T$ for some $\epsilon > 0$. An interesting open problem is to reduce the regret to $N \log T$.

7. REFERENCES

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

[2] P. Rusmevichientong and J. N. Tsitsiklis, "Linearly parameterized bandits," *Mathematics of Operations Research*, vol. 35, no. 2, pp. 395–411, 2010.

[3] L. Li, C. Wei, J. Langford, and R. E. Schapire, "A contextual-bandit approach to personalized news article recommendation," in *Proceeding of International Word Wide Web conference, WWW*, NC, USA, April 2010.

[4] D. Jannach, M. Zanker, A. Felfernig, and G. Friedrich, *Recommender Systems: An Introduction*. Cambridge University Press, 2010.

[5] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 4, pp. 731–745, 2011.

[6] L. Lai, H. E. Gamal, and V. Poor, "Cognitive medium access: Exploration, exploitation, and competition," *IEEE Transactions on Mobile Computing*, vol. 10, no. 2, pp. 239–253, 2011.

[7] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized learning for multiplayer multiarmed bandits," *IEEE Transactions on Information Theory*, vol. 60, no. 4, pp. 2331–2345, 2014.

[8] L. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Transactions on Signal Processing*, vol. 58, no. 11, pp. 5667–5681, 2010.

[9] P. Auer, N. Cesa-Bianchi, Y. F. Robert, and E. Schapire, "The non-stochastic multi-armed bandit problem," *SIAM Journal on Computing*, vol. 32, 2003.

[10] O. Avner, S. Mannor, and O. Shamir, "Decoupling exploration and exploitation in multi-armed bandits," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

[11] M. Lelarge, A. Proutiere, and S. Talebi., "Spectrum bandit optimization," in *Proceedings of IEEE Information Theory Workshop (ITW)*, 2013.

[12] Y. Gai, B. Krishnamachari, and R. Jain, "Combinatorial network optimization with unknown variables: Multi-armed banditswith linear rewards and individual observations," *IEEE/ACM Transactions on Networking*, vol. 20, no. 12, 2012.

[13] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback," in *Proceeding of Conference on Learning Theory, COLT*, Helsinki, Finland, July 2008.

[14] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematics Society*, vol. 58, pp. 527–535, 1952.

[15] M. Hanawal, A.Lesham, and V. Saligrama, "Algorithms for linear bandits on polyhedral sets," *Available at arxiv.com*, 2015.

[16] J.-Y. Audibert, R. Munos, and C. Szepesvári, "Explorationexploitation tradeoff using variance estimates in multi-armed bandits," *Theoretical Computer Science*, vol. 410, p. 18761902, 2009.

[17] Y. Abbasi-Yadkori, D. Pal, and C. Szepesvari, "Improved algorithms for linear stochastic bandits," in *Proceeding of NIPS*, Granada, Spain, Decemeber 2011.

[18] Y. Abbasi-Yadkori, A. Antos, and C. Szepesvári, "Forced-exploration based algorithms for playing in stochastic linear bandits," in *Proceeding COLT workshop on On-line Learning with Limited Feedback*, 2009.

[19] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization*. Athena Scientific, Belmont, Massachusetts, 2008.

[20] P. Auer, Nicholó-Cesa-Bianchi, and P. Fischer, "Finite-time analysis of multiarmed bandit problem trade-offs," *Journal of Machine Learning*, vol. 3, pp. 235–256, 2002.

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

[22] Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári, "Improved algorithms for linear stochastic bandits," in *Proceeding of Advances in Neural Information Processing Systems (NIPS)*, 2011, pp. 2312–2320.