Stochastic Low-Rank Latent Bandits

Adobe Advisor(s): Branislav Kveton, Anup Rao, Zheng Wen *Intern: Subhojyoti Mukherjee*

Abstract

To be written.

1 Introduction

In this paper, we study the problem of recommending the best items to users who are coming sequentially. The learner has access to very less prior information about the users and it has to adapt quickly to the user preferences and suggest the best item to each user. Furthermore, we consider the setting where users are grouped into clusters and within each cluster the users have the same choice of the best item, even though their quality of preference may be different for the best item. These clusters along with the choice of the best item for each user are unknown to the learner. Also, we assume that each user has a single best item preference.

This complex problem can be conceptualized as a low rank stochastic bandit problem where there are K users and L items. The reward matrix, denoted by $\bar{R} \in [0,1]^{K \times L}$, generating the rewards for user, item pair has a low rank structure. The online learning game proceeds as follows, at every timestep t, nature reveals one user (or row) from $\bar{R}$ where user is denoted by i_t. The learner selects one item (or column) from $\bar{R}$, where the item is denoted by j_t. Then the learner receives one noisy feedback $r_t(i_t, j_t) \sim \mathcal{N}(\bar{R}(i_t, j_t)), \sigma^2)$, where $\mathcal{N}$ is a distribution over the entries in $\bar{R}$, σ^2 is variance and $\mathbb{E}[r_t(i_t, j_t)] = \bar{R}(i_t, j_t)$. Then the goal of the learner is to minimize the cumulative regret by quickly identifying the best item j_t^* for each $i_t \in \bar{R}$ where $\bar{R}_{i_t, j_t^*} = \arg\max_{j \in [L]} \{\bar{R}_{i_t, j}\}$.

1.1 Notations, Problem Formulation and Assumptions

We define $[n] = \{1, 2, \ldots, n\}$ and for any two sets A and B, A^B denotes the set of all vectors who take values from A and are indexed by B. Let, $R \in [0,1]^{K \times L}$ denote any matrix, then $R(I, :)$ denote any submatrix of k rows such that $I \in [K]^k$ and similarly $R(:, J)$ denote any submatrix of j columns such that $J \in [L]^j$.

Let $\bar{R}$ be reward matrix of dimension $K \times L$ where K is the number of user or rows and L is the number of arms or columns. Also, let us assume that this matrix $\bar{R}$ has a low rank structure of rank $d << \min\{L, K\}$. Let U and V denote the latent matrices for the users and items, which are not visible to the learner such that,

$$\bar{R} = UV^\intercal \qquad \textbf{s.t.} \qquad U \in [\mathbb{R}^+]^{K \times d}, V \in [0,1]^{L \times d}$$

Furthermore, we put a constraint on V such that, $\forall j \in [L]$, $||V(j, :)||_1 \leq 1$.

Assumption 1. *We assume that there exists d-column base factors, denoted by $V(J^*, :)$, such that all rows of V can be written as a convex combination of $V(J^*, :)$ and the zero vector and $J^* = [d]$. We denote the column factors by $V^* = V(J*, :)$. Therefore, for any $i \in [L]$, it can be represented by*

$$V(i, :) = a_i V(J^*, :),$$

where $\exists a_i \in [0,1]^d$ and $||a_i||_1 \leq 1$.

In this paper, in addition to the noisy setting explained in section 1 we first analyze the proposed algorithm in the easier noise free setting. In the noise free setting, the nature reveals the row i_t, and when the learner selects the column j_t, it observes the mean of the distribution $\bar{R}(i_t, j_t)$.

Assumption 2. *We assume that nature is revealing the user i in $\bar{R}(i,:), \forall i \in [K]$ in a Round-Robin fashion such that at timestep t, nature reveals $i_t = (t \mod K) + 1$.*

The main goal of the learning agent is to minimize the cumulative regret until the end of horizon T. We define the cumulative regret, denoted by $\mathcal{R}_T$ as,

$$\mathcal{R}_T = \sum_{t=1}^{T} \left\{ r_t\left(i_t, j_t^*\right) - r_t\left(i_t, j_t\right) \right\}$$

where, $j_t^* = \arg\max_{j \in [L]} \{\bar{R}(i_t, j)\}$ and j_t be the suggestion of the learner for the i_t -th user. Note that $r_t\left(i_t, j_t^*\right) \sim \mathcal{N}(\bar{R}\left(i_t, j_t^*\right), \sigma^2)$ and $r_t\left(i_t, j_t\right) \sim \mathcal{N}(\bar{R}\left(i_t, j_t\right), \sigma^2)$. Taking expectation over both sides, we can show that,

$$\mathbb{E}[\mathcal{R}_T] = \mathbb{E}\left[\sum_{t=1}^{T} \left\{ r_t\left(i_t, j_t^*\right) - r_t\left(i_t, j_t\right) \right\}\right] = \mathbb{E}\left[\sum_{t=1}^{T} n_{i_t, j_t}\right] \Delta_{i_t, j_t}$$

where, $\Delta_{i_t, j_t} = \bar{R}(i_t, j_t^*) - \bar{R}(i_t, j_t)$ and n_{i_t, j_t} is the number of times the learner has observed the j_t-th item for the i_t-th user. Let, $\Delta = \min_{i \in [K], j \in [L]} \{\Delta_{i,j}\}$ be the minimum gap over all the user, item pair in $\bar{R}$.

1.2 Related Works

In Maillard and Mannor (2014) the authors propose the Latent Bandit model where there are two sets: 1) set of arms denoted by $\mathcal{A}$ and 2) set of types denoted by $\mathcal{B}$ which contains the latent information regarding the arms. The latent information for the arms are modeled such that the set $\mathcal{B}$ is assumed to be partitioned into $|C|$ clusters, indexed by $\mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_C \in \mathcal{C}$ such that the distribution $v_{a,b}, a \in \mathcal{A}, b \in \mathcal{B}_c$ across each cluster is same. Note, that the identity of the cluster is unknown to the learner. At every timestep t, nature selects a type $b_t \in \mathcal{B}_c$ and then the learner selects an arm $a_t \in \mathcal{A}$ and observes a reward $r_t(a, b)$ from the distribution $v_{a,b}$.

Another way to look at this problem is to imagine a matrix of dimension $|A| \times |B|$ where again the rows in $\mathcal{B}$ can be partitioned into $|C|$ clusters, such that the distribution across each of this clusters are same. Now, at every timestep t one of this row is revealed to the learner and it chooses one column such that the $v_{a,b}$ is one of the $\{v_{a,c}\}_{c \in \mathcal{C}}$ and the reward for that arm and the user is revealed to the learner.

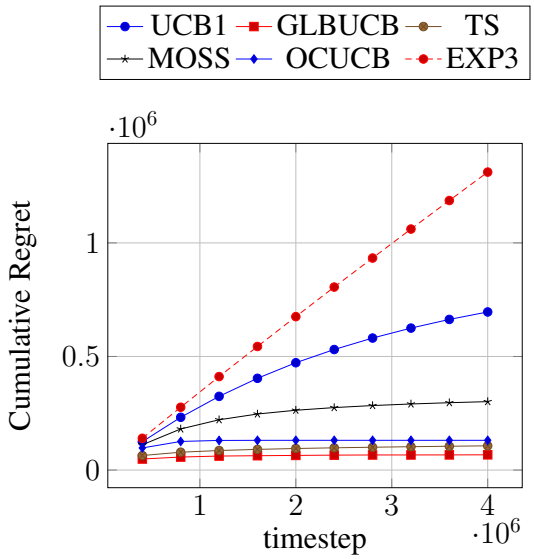
This is actually a much simpler approach than the setting we considered because note that the distributions across each of the clusters $\{v_{a,c}\}_{c \in \mathcal{C}}$ are identical and estimating one cluster distribution will reveal all the information of the users in each cluster.

2 Contributions

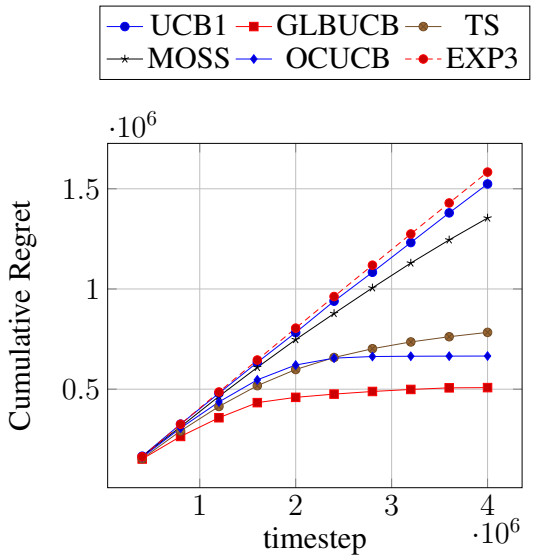
To be written.

3 Proposed Algorithms

4 Experiments



(a) Expt-1: 1024 Users, 16 Bernoulli-distributed arms, Round-Robin, Noisy Setting, Rank 2



(b) Expt-2: 1024 Users, 128 Bernoulli-distributed arms, Round-Robin, Noisy Setting, Rank 3

Figure 1: A comparison of the cumulative regret incurred by the various bandit algorithms.

5 Conclusions and Future Direction

To be written.

References

Maillard, O.-A. and Mannor, S. (2014). Latent bandits.