Generalized 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, L items and the users are coming sequentially. 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 $X_{i_t,j_t} \sim \mathcal{G}(\bar{R}_{i_t,j_t}, \sigma^2)$ from this reward matrix, where $\mathcal{G}$ is a distribution over the entries in $\bar{R}$ and $\mathbb{E}[X_{i_t,j_t}] = \bar{R}_{i_t,j_t}$. 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 $\bar{R}_{i_t,j_t^*} = \arg\max_{j\in[L]}\{\bar{R}_{i_t,j_t}\}$. Also, there exist C clusters, indexed from $\{1,2,\ldots,C\} \in \mathcal{C}$ such that for each $i_t \in U$,

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,

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

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* $||a_i||_1 \leq 1$.

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 $X_{a,b,t}$ 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

3.1 Noise-Free Setting

In the noise-free setting, in addition to Assumption 1 and Assumption 1 we assume two further assumptions.

Assumption 2. *We assume that nature is revealing the* i *of* $\bar{R}(i,:), \forall i \in [K]$ *in a Round-Robin fashion.*

Assumption 3. *We assume that any* d *sets of rows or columns are independent in* $\bar{R}$.

4 Main Results

Lemma 1. *For any arbitrary row* $i \in [K]$,

$$\arg\max_{j\in[L]} U(i,:)V(j,:)^\intercal \leq \arg\max_{j\in[d]} U(i,:)V(j,:)^\intercal.$$

Algorithm 1 Noise-Free GLB

Input: Time horizon T, $Rank(\bar{R}) = d$.
Initialization: Randomly select d columns with uniform probability and $J \leftarrow \{d\}$, $\mathcal{B} \leftarrow \{\emptyset\}$ and $\forall I \in [K], J_0 \in [L], \hat{R}(I, J_0) \leftarrow 0$.
Then, for all the rows $I \in [K]$ observe $\bar{R}(I, J)$ and update $\hat{R}(I, J) \leftarrow \bar{R}(I, J)$.
Explore: Randomly select test column $c \in \mathcal{A} \setminus \{\mathcal{B} \cup J\}$ with uniform probability and set $r_c \leftarrow 0$.
for $t = Kd + 1, .., T$ **do**
  Nature reveals i_t such that $i_t \leftarrow (t \mod K) + 1$ (Round-Robin).
  if $|\mathcal{B}| < L - d$ **then Explore**
    if $r_c < d$ **then**
      Choose c, observe $\bar{R}(i_t, c)$ and $\hat{R}(i_t, c) \leftarrow \bar{R}(i_t, c)$.
      $r_c \leftarrow r_c + 1$
    else
      Column Elimination

        if $\forall i \in [K] : \max_{j \in J} \hat{R}(i, j) \geq \hat{R}(i, c)$ **then**
          $\mathcal{B} \leftarrow \mathcal{B} \cup c$ (Eliminate c)
          Randomly select another test column $c \in \mathcal{A} \setminus \{\mathcal{B} \cup J\}$ with uniform probability.
          $r_c \leftarrow 0$
        else
          $\exists j' \in J : \forall i \in [K] : \max_{j \in J \setminus j' \cup c} \hat{R}(i, j) \geq \hat{R}(i, j')$
          $\mathcal{B} \leftarrow \mathcal{B} \cup j'$ (Eliminate j')
          $J \leftarrow J \setminus j' \cup c$ (Add c as best candidate column)
          $r_c \leftarrow d - K$ (Fully explore new best candidate column)

  else if $|\mathcal{B}| = L - d$ **then Exploit**
    Select column j_t^*, observe $\bar{R}(i_t, j_t^*)$ where $j_t^* \leftarrow \arg\max_{j \in [J]} \hat{R}(i_t, j)$ and $\hat{R}(i_t, , j_t^*) \leftarrow \bar{R}(i_t, , j_t^*)$.

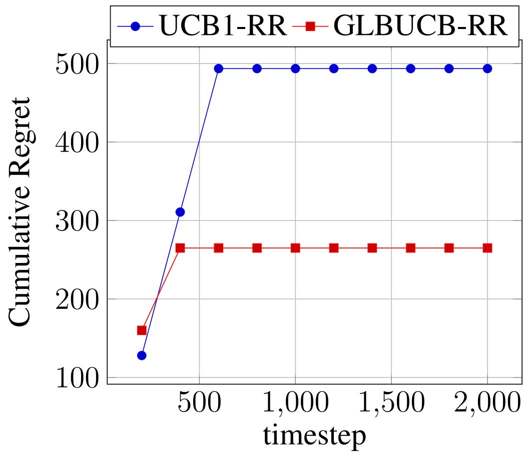
5 Proofs

Proof. Considering any arbitrary row $i \in [K]$, we can show that,

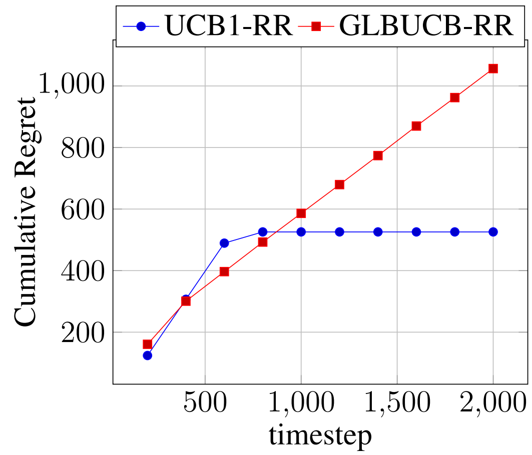
$$
\begin{aligned}
\arg\max_{j \in [L]} U(i, :)V(j, :)^\intercal &= U(i, :)V(j^*(i), :)^\intercal \\
&\overset{(a)}{=} U(i, :) \left(a_{j^*(i)} V(J^*, :)\right)^\intercal \\
&= \sum_{k=1}^{d} a_{j^*(i)}(k) U(i, :)V(j^*(i), :)^\intercal \\
&\leq \arg\max a_{j^*(i)}(k) U(i, :)V(k, :)^\intercal \\
&\leq \arg\max_{k \in [d]} U(i, :)V(k, :)^\intercal,
\end{aligned}
$$

where (a) is from Assumption 1. □

Proof. **Step 1:** We denote the number of times □



(a) Expt-1: 64 Users, 10 Bernoulli-distributed arms, Round-Robin, Noise-Free Setting, Well-Structured Environment



(b) Expt-2: 64 Users, 10 Bernoulli-distributed arms, Round-Robin, Noise-Free Setting, Deformed Environment

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

6 Experiments

7 Conclusions and Future Direction

To be written.

References

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