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

We propose the GLBUCB algorithm based on UCB-Improved, which is an arm elimination algorithm from Auer and Ortner (2010) and is suitable for the stochastic bandit setting. Both these algorithms are phase based column (arm) elimination algorithms where in each phase we select the surviving columns some number of times and then eliminate some sub-optimal columns based on an elimination criteria.

Initialization: The algorithm is initialized with the estimate $\hat{R}_{i,j} = 0, \forall i \in [K], j \in [L]$. In the m-th phase, it denotes the set of surviving columns as $\mathcal{B}_m$. The rows (users) are divided into equivalence classes which are contained in $\mathcal{C}$. N_m denotes the phase length for the m-th phase and each phase length consist of $\gamma|B_m|\ell_m$ timesteps, where $\gamma = f(d) \geq d^2$ is the exploration parameter which is a function of the rank d and ℓ_m is the number of pulls GLBUCB allocates for each of the γ columns.

Equivalence class and initialization: Each equivalence class is denoted as $\mathcal{G}_b = \{i \in [1 + (b-1)d, bd]\}$, where b is indexed from $1, 2, \ldots, \frac{K}{d}$ and these classes are contained in $\mathcal{C}$. At the start of the algorithm γ random columns are chosen for all users in each $\mathcal{G}_b, \forall b \in \left[1, \frac{K}{d}\right]$ and these are contained in $\mathcal{Z}_{\mathcal{G}_b,m}$.

Optimistic greedy sampling: Unlike UCB-Improved, GLBUCB does not pull all surviving columns in the equivalence class for each given user an equal number of times as this wastes a large number of pulls in exploration. Rather, for the entire phase length N_m, it behaves greedily (like UCB1) and selects the column $j_0 \in \arg\max_{j \in \mathcal{B}_m \cap \mathcal{Z}_{\mathcal{G}_b,m}} \left\{\hat{R}(i_t, j) + U_m(\epsilon_m, n_{i_t,j})\right\}$, where i_t is the user revealed by nature and $i_t \in \mathcal{G}_b$. The confidence interval $U_m(\epsilon_m, n_{i_t,j})$ makes sure that sufficient exploration is conducted amongst the columns in $\mathcal{Z}_{\mathcal{G}_b,m}$.

Structured Column Elimination: In the column elimination sub-module GLBUCB eliminates a sub-optimal column by making sure that it is not one of the d-best columns. Moreover, in the column elimination sub-module, the confidence interval $U_m(\epsilon_m, n_m)$ helps in eliminating a sub-optimal column with high probability in the noisy setting.

Reset Parameters: The reset parameters sub-module, can be divided into two parts:

- **Information share:** GLBUCB reconstructs the equivalence classes such that the each $\mathcal{G}_b, \forall b \in \left[1, \frac{K}{d}\right]$ contains the d best performing columns amongst all the users $i \in [K]$. This is achieved by first selecting the column $j_0 \in \arg\max_{j \in \mathcal{Z}_{\mathcal{G}_b,m}} \left\{\hat{R}(i,j) + U_m(\epsilon_m, n_{i,j})\right\}, \forall i \in [K], i \in \mathcal{G}_b$ and then selecting the d most frequent such columns. These d-best columns are included in $\mathcal{Z}_{\mathcal{G}_b,m+1}, \forall b \in \left[1, \frac{K}{d}\right]$ and then for each $\mathcal{G}_b$ the remaining $\gamma - d$ columns are selected uniform randomly from $\mathcal{B}_m \cap \mathcal{Z}_{\mathcal{G}_b,m}$.
- **Increase exploration:** GLBUCB increases the exploration bonus for the next phase so that more exploration is conducted for the surviving columns.

Finally, if the algorithm has eliminated $L - d$ columns, for the user i_t revealed by the nature, it always selects the column j_t^*, where $j_t^* \leftarrow \arg\max_{j \in [\mathcal{B}_m]} \hat{R}(i_t, j)$. The pseudo-code of this is shown in Algorithm 1.

4 Main Results

4.1 Regret Bound of GLBUCB

5 Proofs

Proof. **Step 0. (Outline):** We separate the proof into two larger sub-modules. In the first sub-module, we show that there exist a phase $m_{i,j}$ for a user $i \in \mathcal{G}_b$ such that the d-best items are in $\mathcal{Z}_{\mathcal{G}_b,m_{i,j}}$ with a very high probability. In

Algorithm 1 GLBUCB

Input: Time horizon T, $Rank(\bar{R}) = d$.

Explore Parameters: $\gamma \geq d^2, \alpha \geq \frac{1}{2}, \psi \geq 1$.

Definition: $U_m(\epsilon_m, n_{i,j}) = \sqrt{\frac{\alpha \log(\psi T \epsilon_m^2)}{2n_{i,j}}}$

Initialization: $\hat{R}(i,j) \leftarrow 0, \forall i \in [K], \forall j \in [L]$, $m = 0$, $\mathcal{B}_0 \leftarrow \mathcal{A}$, $\epsilon_0 = 1$, $\ell_0 = \frac{2\log(\psi T \epsilon_0^2)}{\epsilon_0}$ and $N_0 = K\gamma\ell_0$.

for each $b \in \left[1, \frac{K}{d}\right]$ **do** ▷ Create equivalence class $\mathcal{C}$

  $\mathcal{G}_b \leftarrow \{i \in [1 + (b-1)d, bd]\}$ and $\mathcal{C} \leftarrow \mathcal{C} \cup \mathcal{G}_b$.

$Z_{\mathcal{G}_b,0} \leftarrow$ Select γ random columns for each equivalence class $\mathcal{G}_b \in \mathcal{C}, \forall b \in \left[1, \frac{K}{d}\right]$ and ensure $\bigcup_{b=1}^{\frac{K}{d}} Z_{\mathcal{G}_b,0} = [L]$.

for $t = 1, .., T$ **do**

  Nature reveals i_t s.t. $i_t \leftarrow (t \mod K) + 1$ ▷ Round-Robin

  if $|\mathcal{B}_m| > d$ **then**

    if $t \leq N_m$ **then** ▷ Till end of Phase do UCB1 on $\mathcal{Z}_{\mathcal{G}_b,m}$

      Choose $j_0 \in \arg\max_{j \in \mathcal{Z}_{\mathcal{G}_b,m}} \left\{\hat{R}(i_t, j) + U_m(\epsilon_m, n_{i_t,j})\right\}$, where $i_t \in \mathcal{G}_b$.

    else ▷ End of phase, do elimination and reset parameters

      Structured Column Elimination

        for each $\mathcal{G}_b \in \mathcal{C}$ **do**

          while $\exists j \in \mathcal{B}_m$ **s.t.** $\forall i \in \mathcal{G}_b : \hat{R}(i,j) + U_m(\epsilon_m, n_{i,j}) < \max_{j' \in \mathcal{B}_m \setminus j} \left\{\hat{R}(i,j') - U_m(\epsilon_m, n_{i,j'})\right\}$ **do**

            $\mathcal{B}_m \leftarrow \mathcal{B}_m \setminus \{j\}$.

      Reset Parameters

        for each $i \in [K]$ **do** ▷ Find best d arms

          $\mathcal{D}_m(i) \leftarrow \arg\max_{j \in \mathcal{Z}_{\mathcal{G}_b,m}} \left\{\hat{R}(i,j) + U_m(\epsilon_m, n_{i,j})\right\}, \forall j \in \mathcal{B}_m$

        for each $\mathcal{G}_b \in \mathcal{C}$ **do**

          $\mathcal{Z}_{\mathcal{G}_b,m+1} \leftarrow d$ most frequent columns in $\mathcal{D}_m$ and select $\gamma - d$ different columns uniform randomly from $\mathcal{B}_m \cap \mathcal{Z}_{\mathcal{G}_b,m}$ and ensure $\bigcup_{b=1}^{\frac{K}{d}} Z_{\mathcal{G}_b,m+1} = \{B_m\}$.

        $\epsilon_{m+1} \leftarrow \frac{\epsilon_m}{2}$ and $\ell_{m+1} \leftarrow \frac{2\log(\psi T \epsilon_{m+1}^2)}{\epsilon_{m+1}}$

        $N_{m+1} \leftarrow t + K\gamma\ell_{m+1}$ and $m \leftarrow m + 1$.

  else ▷ Till T do UCB1 on remaining d arms

    Select column $j_t^* \in \arg\max_{j \in \mathcal{B}_m} \left\{\hat{R}(i_t, j) + U_m(\epsilon_m, n_{i_t,j})\right\}$

the second sub-module, we show that in such a phase $m_{i,j}$, there exist a $\mathcal{G}_b$ such that $J^* \in \mathcal{G}_b$ and for all $i \in \mathcal{G}_b$ a sub-optimal item j is eliminated with a very high probability.

Step 1.(Define some notations): In this proof, we define the confidence interval for the (i,j)-th user-item pair as $S_{i,j} = \sqrt{\frac{\alpha \log(\psi T \epsilon_m^2)}{n_{i,j}}}$. Let J^* denote the set of d-best items. The phase numbers are denoted by $m = 0, 1, \ldots, M$

where $M = \frac{1}{2}\log_2 \frac{T}{e}$. We also define $\mathcal{A}' = \left\{i \in [K], j \in \mathcal{A} : \Delta_{i,j} \geq \sqrt{\frac{e}{T}}\right\}$.

Step 2.(Define a phase p_{J^*}): We define a phase p_{J^*} such that all $\mathcal{Z}_{\mathcal{G}_b,m}, \forall b \in \left[1, \frac{K}{d}\right]$ contains J^*.

Step 3.(Frequency of UCBs from J^* in D_m): In the m-th phase, for the best-d column selection, GLBUCB selects $\arg\max_{j \in \mathcal{Z}_{\mathcal{G}_b,m}} \left\{\hat{R}(i,j) + U_m(\epsilon_m, n_{i,j})\right\}$ for each $i \in [K]$ and $i \in \mathcal{Z}_{\mathcal{G}_b,m}$. Let, z_j^* be the count of the total number of times the UCB from $j \in J^*$ is in D_m. Let, z_j be the count of the total number of times the UCB from j

Step 3.(Regret for selecting sub-optimal item j after p_{J^*}-th phase): Note, that by construction of GLBUCB, for any phase m, $\bigcup_{b=1}^{\frac{K}{d}} Z_{\mathcal{G}_b,m+1} = \{B_m\}$. For all $\mathcal{Z}_{\mathcal{G}_b}$ to contain J^*, $\mathcal{D}_{p_{J^*}}(i)$ must contain $j^*, \forall i \in [K]$. So, if the following three conditions hold, then

Step 3.(Define a stopping phase for item j and $i \in \mathcal{G}_b$ as $m_{i,j}$): We define a stopping phase $m_{i,j}$ for a sub-optimal item $j \in \mathcal{A}'$ as the first phase after which the item j is eliminated,

$$m_{i,j} = \min\left\{m : \sqrt{\alpha\epsilon_m} < \frac{\Delta_{i,j}}{4}\right\}$$

Note, that this item j is no longer selected for any user $i \in [K]$.

Step 4.(Regret for sub-optimal item j being selected after m_j-th phase): Note, that on or after the $m_{i,g}$-th phase a sub-optimal arm $j \in \mathcal{A}'$ is eliminated if these four conditions hold for all $i \in \mathcal{G}_b$,

$$\hat{R}(i,j) < \bar{R}(i,j) + S_{i,j}, \qquad \hat{R}(i,j^*) > \bar{R}(i,j^*) - S_{i,j^*}, \qquad S_{i,j} > S_{i,j^*}, \qquad n_{i,j} \geq \ell_{m_{i,j}} \tag{1}$$

Moreover, in the $m_{i,j}$-th phase for an $i \in \mathcal{G}_b$ if $n_{i,j} \geq \ell_{m_{i,j}} = \dfrac{\log(\psi T \epsilon_{m_{i,j}}^2)}{2\epsilon_{m_{i,j}}}$ then we can show that,

$$S_{i,j} = \sqrt{\frac{\alpha\log(\psi T\epsilon_{m_{i,j}}^2)}{2n_{i,j}}} \leq \sqrt{\frac{\alpha\log(\psi T\epsilon_{m_{i,j}}^2)}{2\ell_{m_{i,j}}}} \leq \sqrt{\alpha\epsilon_{m_j}\frac{\log(\psi T\epsilon_{m_{i,j}}^2)}{\log(\psi T\epsilon_{m_j}^2)}} \leq \frac{\Delta_{i,j}}{4}.$$

If the four conditions in equation 1 hold for all $i \in \mathcal{G}_b$ then we can show that in the m_j-th phase,

$$\begin{aligned}
\hat{R}(i,j) + S_{i,j} &\leq \bar{R}(i,j) + 4S_{i,j} - 2S_{i,j} \\
&\leq \bar{R}(i,j) + \Delta_{i,j} - 2S_{i,j} \\
&\leq \bar{R}(i,j^*) - 2S_{i,j^*} \\
&\leq \hat{R}(i,j^*) - S_{i,j^*}
\end{aligned}$$

Hence, the sub-optimal item j is eliminated in the m_j-th phase. Therefore, to bound the number of pulls of the sub-optimal item j, we need to bound the probability of the complementary of the four events in equation **??**.

For the first event in equation 1, using Chernoff-Hoeffding bound we can upper bound the probability of the complementary of that event by,

$$\sum_{i=1}^{d}\sum_{m=0}^{m_{i,j}}\sum_{n=1}^{\ell_m}\mathbb{P}\Big\{\frac{\sum_{s=1}^{n} r_s(i,j)}{n} \geq \bar{R}(i,j)+\sqrt{\frac{\alpha\log(\psi T\epsilon_{m_{i,j}}^2)}{2n}}\Big\} \leq \sum_{i=1}^{d}\sum_{m=0}^{m_{i,j}}\sum_{n=1}^{\ell_m}\exp\left(-2\left(\sqrt{\frac{\alpha\log(\psi T\epsilon_m^2)}{2n}}\right)^2 n\right)$$
$$\leq \sum_{i=1}^{d}\sum_{m=0}^{m_{i,g}}\sum_{n=1}^{\ell_m}\exp\left(-2\frac{\alpha\log(\psi T\epsilon_m^2)}{2n}n\right)$$
$$\leq \sum_{i=1}^{d}\sum_{m=0}^{m_{i,j}}\sum_{n=1}^{\ell_m}\frac{1}{(\psi T\epsilon_m^2)^{\alpha}}$$
$$\leq \sum_{i=1}^{d}\sum_{i=1}^{d}\sum_{m=0}^{m_{i,j}}\frac{\log(\psi T\epsilon_m^2)}{2\epsilon_m}\frac{1}{(\psi T\epsilon_m^2)^{\alpha}}$$
$$\leq \sum_{i=1}^{d}\frac{\log(\psi T\sum_{m=0}^{m_{i,j}}\epsilon_m^2)}{2(\psi T)^{\alpha}}\sum_{m=0}^{m_{i,j}}\frac{1}{\epsilon_m^{2\alpha+1}}$$
$$\leq \sum_{i=1}^{d}\frac{\log(\psi T)}{2(\psi T)^{\alpha}}\sum_{m=0}^{m_{i,j}}\frac{1}{\epsilon_m^{2\alpha+1}}.$$

Similarly, for the second event in equation 1, we can bound the probability of its complementary event by,

$$\sum_{i=1}^{d}\sum_{m=0}^{m_{i,j}}\sum_{n=1}^{\ell_m}\mathbb{P}\Big\{\frac{\sum_{s=1}^{n} r_s(i,j^*)}{n} \leq \bar{R}(i,j^*)-\sqrt{\frac{\alpha\log(\psi T\epsilon_m^2)}{2n}}\Big\} \leq \sum_{i=1}^{d}\sum_{m=0}^{m_{i,j}}\sum_{n=1}^{\ell_m}\exp\left(-2\left(\sqrt{\frac{\alpha\log(\psi T\epsilon_m^2)}{2n}}\right)^2 n\right)$$
$$\leq \sum_{i=1}^{d}\frac{\log(\psi T)}{2(\psi T)^{\alpha}}\sum_{m=0}^{m_{i,j}}\frac{1}{\epsilon_m^{2\alpha+1}}.$$

Also, for the third event in equation 1, we can bound the probability of its complementary event by,

$$\sum_{i=1}^{d}\sum_{m=0}^{m_{i,j}}\mathbb{P}\{S_{i,j} < S_{i,j^*}\} \leq \sum_{i=1}^{d}\sum_{m=0}^{m_{i,j}}\mathbb{P}\{\hat{R}(i,j)+S_{i,j} > \hat{R}(i,j^*)+S_{i,j^*}\}$$

Following, the argument of Auer et al. (2002) we can show that the event $\hat{R}(i,j)+S_{i,j} > \hat{R}_{i,j^*}+S_{i,j^*}$ is possible only when the following three events occur for each $i \in \mathcal{G}_b$,

$$\hat{R}(i,j^*) \leq \hat{R}(i,j^*)-S_{i,j^*}, \qquad \hat{R}(i,j) \geq \hat{R}(i,j)+S_{i,j}, \qquad \bar{R}(i,j^*)-\bar{R}(i,j) < 2S_{i,j}.$$

However, the third event will not happen with high probability for $n_{i,j} \geq \ell_{m_j}$. Proceeding as before, we can show that the probability of the remaining two events is bounded by,

$$\sum_{i=1}^{d}\sum_{m=0}^{m_{i,j}}\mathbb{P}\{S_{i,j}<S_{i,j^*}\}\leq\sum_{i=1}^{d}\sum_{m=0}^{m_{i,j}}\sum_{n=1}^{\ell_m}\sum_{q=1}^{\ell_m}\mathbb{P}\left\{\frac{\sum_{s=1}^{n}r_s(i,j)}{n}+\sqrt{\frac{\alpha\log(\psi T\epsilon_m^2)}{2n}}>\frac{\sum_{s=1}^{q}r_s(i,j^*)}{q}+\sqrt{\frac{\alpha\log(\psi T\epsilon_m^2)}{2q}}\right\}$$

$$\leq\sum_{i=1}^{d}\sum_{m=0}^{m_{i,j}}\sum_{n=1}^{\ell_m}\mathbb{P}\left\{\frac{\sum_{s=1}^{n}r_s(i,j)}{n}\geq\bar{R}(i,j)+\sqrt{\frac{\alpha\log(\psi T\epsilon_m^2)}{2n}}\right\}$$

$$+\sum_{i=1}^{d}\sum_{m=0}^{m_{i,j}}\sum_{q=1}^{\ell_m}\mathbb{P}\left\{\frac{\sum_{s=1}^{q}r_s(i,j^*)}{q}\leq\bar{R}(i,j^*)-\sqrt{\frac{\alpha\log(\psi T\epsilon_m^2)}{2q}}\right\}$$

$$\leq\sum_{i=1}^{d}\sum_{m=0}^{m_{i,j}}\sum_{n=1}^{\ell_m}\exp\left(-2\frac{\alpha\log(\psi T\epsilon_m^2)}{2n}n\right)+\sum_{i=1}^{d}\sum_{m=0}^{m_{i,g}}\sum_{q=1}^{\ell_m}\exp\left(-2\frac{\alpha\log(\psi T\epsilon_m^2)}{2q}q\right)$$

$$\leq\sum_{i=1}^{d}\frac{\log(\psi T)}{2(\psi T)^{\alpha}}\sum_{m=0}^{m_{i,g}}\frac{1}{\epsilon_m^{2\alpha+1}}$$

Finally, for the fourth event in equation 1 we can show that for each $i\in\mathcal{G}_b$,

$$\sum_{i=1}^{d}\sum_{m=0}^{m_{i,j}}\mathbb{P}\{n_{i,j}<\ell_{m_{i,j}}\}\leq\sum_{m=0}^{m_{i,j}}\mathbb{P}\{\hat{\mu}_{i,g}+S_{i,g}<\hat{\mu}_{i*,g}+S_{i*,g}\}$$

$$\leq\sum_{i=1}^{d}\sum_{m=0}^{m_{i,g}}\sum_{n=1}^{\ell_m}\sum_{q=1}^{\ell_m}\mathbb{P}\left\{\frac{\sum_{s=1}^{n}r_s(i,j)}{n}+\sqrt{\frac{\alpha\log(\psi T\epsilon_m^2)}{2n}}<\frac{\sum_{s=1}^{q}r_s(i,j^*)}{q}+\sqrt{\frac{\alpha\log(\psi T\epsilon_m^2)}{2q}}\right\}$$

$$\leq\sum_{i=1}^{d}\frac{\log(\psi T)}{(\psi T)^{\alpha}}\sum_{m=0}^{m_{i,j}}\frac{1}{\epsilon_m^{2\alpha+1}}.$$

Combining the above four cases we can bound the probability that a sub-optimal arm i will no longer be pulled on or after the $m_{i,g}$-th phase by,

$$\sum_{i=1}^{d}\frac{4\log(\psi T)}{(\psi T)^{\alpha}}\sum_{m=0}^{m_{i,j}}\frac{1}{\epsilon_m^{2\alpha+1}}\leq\frac{4\log(\psi T)}{(\psi T)^{\alpha}}\sum_{m=0}^{M}\left(\frac{1}{\epsilon_m}\right)^{2\alpha+1}$$

$$\overset{(a)}{\leq}\sum_{i=1}^{d}\frac{4\log(\psi T)}{(\psi T)^{\alpha}}\left(\frac{2(2^M-1)}{(2)-1}\right)^{2\alpha+1}$$

$$\overset{(b)}{\leq}\sum_{i=1}^{d}\frac{4\log(\psi T)}{(\psi T)^{\alpha}}\left(2\sqrt{T}\right)^{2\alpha+1}=\sum_{i=1}^{d}\frac{2^{2\alpha+3}\sqrt{T}\log(\psi T)}{(\psi)^{\alpha}}.$$

Here, in (a) we use the standard geometric progression formula and in (b) we substitute the value of $M=\frac{1}{2}\log_2\frac{T}{e}$. Bounding this trivially by $T\Delta_{i,j}$ for each item $j\in\mathcal{A}'$ we get the regret suffered for all items $j\in\mathcal{A}'$ and for each $i\in\mathcal{G}_b$ after the $m_{i,j}$-th phase as,

$$\sum_{i=1}^{d}\sum_{j\in\mathcal{A}'}\left(\frac{2^{2\alpha+3}\sqrt{T}\log(\psi T)}{(\psi)^{\alpha}}\right)=\sum_{i=1}^{d}\sum_{j\in\mathcal{A}'}\left(\frac{2^{2\alpha+3}T^{\frac{3}{2}}\Delta_{i,j}\log(\psi T)}{(\psi)^{\alpha}}\right)=\sum_{i=1}^{d}\sum_{j\in\mathcal{A}'}\left(\frac{2^{2\alpha+3}\Delta_{i,j}\log(\psi T)}{(\psi T^{-\frac{3}{2\alpha}})^{\alpha}}\right).$$

Step 6.(Regret for pulling the sub-optimal item j on or before $m_{i,j}$-th phase): Either a sub-optimal item j gets pulled $\ell_{m_{i,j}}$ number of times till the $m_{i,j}$-th phase or after that the probability of it getting pulled is exponentially low (as shown in **step 4**). Hence, the number of times a sub-optimal item j is pulled till the $m_{i,j}$-th phase is given by,

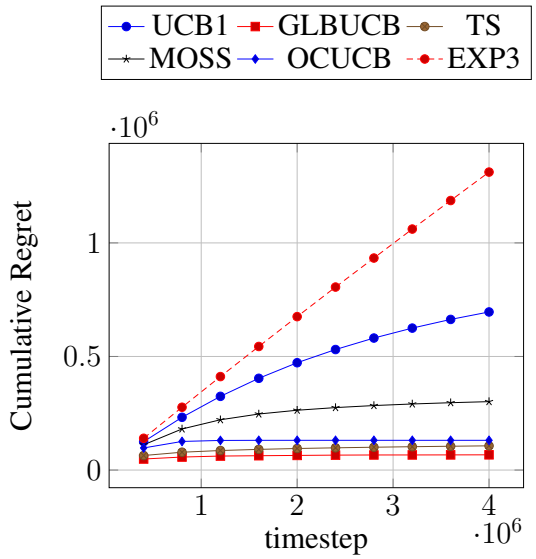
$$n_{i,j}<\ell_{m_{i,j}}=\left\lceil\frac{\log(\psi T\epsilon_{m_{i,j}}^{2})}{2\epsilon_{m_{i,j}}}\right\rceil$$

Hence, considering each item $j\in\mathcal{A}'$ and each $i\in\mathcal{G}_b$ the total regret is bounded by,

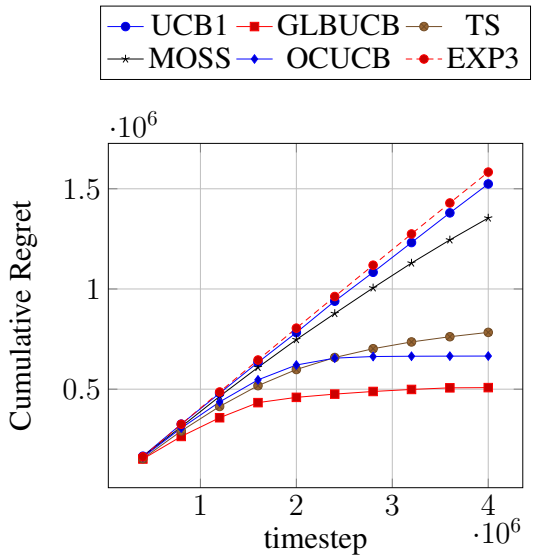
$$\sum_{i=1}^{d}\sum_{i\in\mathcal{A}'}\Delta_{i,j}\left\lceil\frac{\log(\psi T\epsilon_{m_{i,g}}^{2})}{2\epsilon_{m_{i,g}}}\right\rceil<\sum_{i=1}^{d}\sum_{i\in\mathcal{A}'}\Delta_{i,j}\left[1+\frac{\log(\psi T\epsilon_{m_{i,j}}^{2})}{2\epsilon_{m_{i,j}}}\right]\leq\sum_{i=1}^{d}\sum_{i\in\mathcal{A}'}\Delta_{i,j}\left[1+\frac{8\log(\psi T(\Delta_{i,j})^{4})}{(\Delta_{i,j})^{2}}\right].$$

□

6 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.

7 Conclusions and Future Direction

To be written.

References

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256.

Auer, P. and Ortner, R. (2010). Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65.

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