

Combinatorial Categorized Bandits with Expert Rankings (Supplementary Material)

Sayak Ray Chowdhury*¹

Gaurav Sinha*¹

Nagarajan Natarajan¹

Amit Sharma¹

¹Microsoft Research, Bengaluru, India

A PROOFS

A.1 REGRET UPPER BOUND OF ALGORITHM 1

We start by recalling that $\Delta_{z_t} := f^{\mu^*}(z^*) - f^{\mu^*}(z_t)$ denotes the regret or *gap* of selecting allocation z_t at round t instead of the optimal allocation z^* . Now, define the event

$$\mathcal{E}_t = \left\{ \Delta_{z_t} \leq 2 \sum_{(i,j): z_i^* < j \leq z_{t,i}} \sqrt{\frac{1.5 \log T}{T_{t-1}(a_{i,j})}}, \Delta_{z_t} > 0 \right\}.$$

Also, define $\hat{R}_T = \sum_{t=MN+1}^T \Delta_{z_t} \mathbb{1}\{\mathcal{E}_t\}$. Then, from Kveton et al. [2015, Lemma 1], it holds that

$$R_T \leq \mathbb{E} \left[\hat{R}_T \right] + (1 + \pi^2/3) KMN. \quad (1)$$

Now, let us consider two sequence of constants $(\alpha_i)_{i \geq 1}$ and $(\beta_i)_{i \geq 0}$ such that $\beta_0 = 1$, $\alpha_i > \alpha_j$, $\beta_i > \beta_j$ for all $i > j$, $\lim_{i \rightarrow \infty} \alpha_i = \lim_{i \rightarrow \infty} \beta_i = 0$ and $\sum_{i \geq 1} \frac{\beta_{i-1} - \beta_i}{\sqrt{\alpha_i}} \leq 1/\sqrt{6}$.

Let A_t denote the subset of items induced by the allocation z_t chosen at round t . In other words, an item $a_{i,j} \in A_t$ if $z_{t,i} \geq j$. Similarly, let A^* denote the corresponding subset induced by the optimal allocation z^* . Let $\tilde{A}_t = A_t \setminus A^*$ and $m_{i,t} = \frac{\alpha_i K^2 \log T}{\Delta_{z_t}^2}$. Now, similar to Kveton et al. [2015], define the series of mutually exclusive events $(G_{i,t})_{i \geq 1}$, where $G_{i,t}$ denotes the event that at least $\beta_i K$ items in \tilde{A}_t were observed at most $m_{i,t}$ times and for all $j < i$, less than $\beta_1 K$ items in \tilde{A}_t were observed at most $m_{i-1,t}$ times. Then, under the event \mathcal{F}_t , it holds that the event $\{\bigcup_{i \geq 1} G_{i,t}\}$ happens [Kveton et al., 2015, Lemma 3]. Hence, we have

$$\hat{R}_T = \sum_{l=1}^{\infty} \sum_{t=MN+1}^T \Delta_{z_t} \mathbb{1}\{G_{l,t}, \Delta_{z_t} > 0\}.$$

Now for any item $a_{i,j}$, let us define the events

$$F_{a_{i,j},l,t} = \{z_i^* < j \leq z_{t,i}, T_{t-1}(a_{i,j}) \leq m_{l,t}\}, \quad G_{a_{i,j},l,t} = G_{l,t} \cap F_{a_{i,j},l,t}.$$

Let us now define the following events:

$$F_{a_{i,j},l,t}^k = \{z_i^* < j+k = z_{t,i}, T_{t-1}(a_{i,j+k}) \leq m_{l,t}\}, k \geq 0.$$

*Equal contribution

Note that because of the ordered structure. if $a_{i,j}$ has only been observed a certain number of times, then a_{i,j_k} would be observed less than or equal number of times i.e., $T_{t-1}(a_{i,j+k}) \leq T_{t-1}(a_{i,j})$, which, in turn, implies that

$$\tilde{F}_{a_{i,j},l,t}^k := \{z_i^* < j+k = z_{t,i}, T_{t-1}(a_{i,j}) \leq m_{l,t}\} \subseteq F_{a_{i,j},l,t}^k.$$

It turns out that $F_{a_{i,j},l,t} = \bigcup_{k=0}^{M-j} \tilde{F}_{a_{i,j},l,t}^k$, which in turn implies $F_{a_{i,j},l,t} \subseteq \bigcup_{k=0}^{M-j} F_{a_{i,j},l,t}^k =: H_{a_{i,j},l,t}$. This implies that $\bigcup_{j=1}^M \{G_{l,t} \cap F_{a_{i,j},l,t}\} \subseteq \bigcup_{j=1}^M \{G_{l,t} \cap H_{a_{i,j},l,t}\}$. Now observe that $H_{a_{i,1},l,t} \supseteq H_{a_{i,2},l,t} \supseteq \dots \supseteq H_{a_{i,M},l,t}$, implying that the RHS of the above is a union over decreasing sets, and hence it holds that $\bigcup_{j=1}^M \{G_{l,t} \cap F_{a_{i,j},l,t}\} \subseteq \{G_{l,t} \cap H_{a_{i,1},l,t}\}$. This further implies

$$\bigcup_{j=1}^M G_{a_{i,j},l,t} = \bigcup_{j=1}^M \{G_{l,t} \cap F_{a_{i,j},l,t}\} \subseteq \bigcup_{k=0}^{M-1} \{G_{l,t} \cap F_{a_{i,1},l,t}^k\} = \bigcup_{j=1}^M \{G_{l,t} \cap \{z_i^* < j = z_{t,i}, T_{t-1}(a_{i,j}) \leq m_{l,t}\}\}$$

Then, it holds that

$$\mathbb{1}\{G_{l,t}, \Delta_{z_t} > 0\} \leq \frac{1}{\beta_l K} \sum_{(i,j): z_i^* < j} \mathbb{1}\{z_{t,i} = j, T_{t-1}(a_{i,j}) \leq m_{l,t}, \Delta_{z_t} > 0\}.$$

Therefore, we can bound \hat{R}_T as

$$\hat{R}_T \leq \sum_{(i,j): z_i^* < j} \sum_{l=1}^{\infty} \sum_{t=MN+1}^T \mathbb{1}\{z_{t,i} = j, T_{t-1}(a_{i,j}) \leq m_{l,t}, \Delta_{z_t} > 0\} \frac{\Delta_{z_t}}{\beta_l K},$$

Now let each item $a_{i,j}$, which are not included in the optimal allocation z^* , be contained in $N_{i,j}$ suboptimal allocations z and $\Delta_{i,j,1} \geq \Delta_{i,j,2} \geq \dots \geq \Delta_{i,j,N_{i,j}}$ be the gaps of these solutions. Then, we have

$$\begin{aligned} \hat{R}_T &\leq \sum_{(i,j): z_i^* < j} \sum_{l=1}^{\infty} \sum_{t=MN+1}^T \sum_{k=1}^{N_{i,j}} \mathbb{1}\{z_{t,i} = j, T_{t-1}(a_{i,j}) \leq \frac{\alpha_l K^2 \log T}{\Delta_{i,j,k}^2}, \Delta_{z_t} = \Delta_{i,j,k}\} \frac{\Delta_{i,j,k}}{\beta_l K} \\ &\leq \sum_{(i,j): z_i^* < j} \sum_{l=1}^{\infty} \frac{\alpha_l K \log T}{\beta_l} \left(\frac{1}{\Delta_{i,j,1}} + \sum_{k=2}^{N_{i,j}} \Delta_{i,j,k} \left(\frac{1}{\Delta_{i,j,k}^2} - \frac{1}{\Delta_{i,j,k-1}^2} \right) \right), \end{aligned}$$

where the last term is the solution of the optimization problem:

$$\max_{(z_1, \dots, z_t)} \sum_{t=1}^T \sum_{k=1}^{N_{i,j}} \mathbb{1}\{z_{t,i} = j > z_i^*, T_{t-1}(a_{i,j}) \leq \frac{\alpha_l K^2 \log T}{\Delta_{i,j,k}^2}, \Delta_{z_t} = \Delta_{i,j,k}\} \frac{\Delta_{i,j,k}}{\beta_l K}.$$

Then, similar to Kveton et al. [2015], we obtain

$$\hat{R}_T \leq \sum_{(i,j): z_i^* < j} \sum_{l=1}^{\infty} \frac{2\alpha_l K \log T}{\beta_l \Delta_{i,j,N_{i,j}}} \leq \sum_{(i,j): z_i^* < j} \frac{534K \log T}{\Delta_{i,j}},$$

where $\Delta_{i,j}$ denotes the minimum gap as defined in main paper, and is given by $\Delta_{i,j,N_{i,j}}$. Now, from (1), we can complete the proof.

A.2 PROBABILITY OF ERROR OF ALGORITHM 2

As a first step we define an event η which helps us in the rest of the proof. Note that Bubeck et al. [2013] also defined a similar event ξ in their proof of Theorem 1, but our event is different from theirs. In η we only consider the $|\Phi|$ ($\leq 2N$) sized subset of phases where some item from the boundary set Φ is accepted or rejected, whereas Bubeck et al. [2013] considered all the $MN - 1$ phases. Let $k_1 \leq k_2 \leq \dots \leq k_{|\Phi|}$ be the phases where items in Φ were accepted or rejected. Under this notation $H_{\Phi} = \max_{1 \leq i \leq |\Phi|} \frac{MN+1-k_i}{\Delta_{[MN+1-k_i]}}^2$. Consider the event η defined by

$$\{\forall i \in \{1, \dots, MN\}, \forall k \in \{k_1, \dots, k_{|\Phi|}\}, \left| \frac{1}{T_k} \sum_{s=1}^{T_k} X_{i,s} - \mu_i \right| \leq \frac{1}{4} \Delta_{[MN+1-k]}\}$$

Note that by abuse of notation, we have renamed our items as $1, \dots, MN$ above. Also $X_{i,s}$ denotes the bernoulli reward received for item i in its s^{th} pull so far. Recall that T_k was defined as $\frac{T-MN}{\log(MN)(MN+1-k)}$. By Hoeffding's inequality and union bound, we bound the probability of the complement event $\bar{\eta}$ as

$$\begin{aligned} \mathbb{P}(\bar{\eta}) &\leq \sum_{i=1}^{MN} \sum_{j=1}^{|\Phi|} \mathbb{P} \left(\left| \frac{1}{T_{k_j}} \sum_{s=1}^{T_{k_j}} X_{i,s} - \mu_i \right| > \frac{1}{4} \Delta_{[MN+1-k_j]} \right) \\ &\leq \sum_{i=1}^{MN} \sum_{j=1}^{|\Phi|} 2 \exp(-2T_{k_j} (\Delta_{[MN+1-k_j]}/4)^2) \\ &\leq 2MN|\Phi| \exp \left(-\frac{T-MN}{8\log(MN)H_\Phi} \right) \end{aligned}$$

Next, we show that assuming the event η , the algorithm does not make any error. The proof of this part is very similar to the proof of Theorem 1 in Bubeck et al. [2013]. The main difference is that in our case since we always accept items (items) that are in the top of some list and reject items that are in the bottom of some list, we do not make any error until we reach an item in the boundary i.e. Φ . We will claim that the event η prevents these errors from happening. This is done by induction on the phases where some boundary item is accepted or rejected i.e. phases $k_1, \dots, k_{|\Phi|}$. Since we only need to argue about correctness for these phases, we defined η only for $k \in \{k_1, \dots, k_{|\Phi|}\}$. Note that this is the critical difference between our event η and the corresponding event ξ defined in proof of Theorem 1 in Bubeck et al. [2013]). Define $k_0 = 0$ and consider $j \geq 1$. Using an induction approach we assume that no errors have happened till phase k_{j-1} . As explained above the next error can only occur at an item in boundary and therefore has to occur at phase k_j . We show that under event η , this cannot happen. We will need the following observation.

Observation A.1. *Let e be the first item to be erroneously accepted or rejected. As mentioned above $e \in \Phi$. From Algorithm 2, we know that there is some active item e' in the same list as e with the highest empirical gap $\hat{\Delta}_{e'}$ among all the active items. By the design of our algorithm, if e was accepted it is the top item of its list which also contains the active element $e' \Rightarrow \mu_{e'} \leq \mu_e$. Since e was erroneously accepted, it does not belong to the top K items $\Rightarrow e'$ also does not belong to the top K items. Similarly if e was erroneously rejected (i.e. it belongs to top K items), it is the bottom of its list which contains the active element $e' \Rightarrow \mu_e \leq \mu_{e'}$. Thus e' also belongs to top K items.*

Event η implies that at the end of stage k_j , empirical means of rewards of all items are within $\frac{1}{4}\Delta_{[MN+1-k_j]}$ of their true reward means. Let $A_{k_j} = \{a_1, \dots, a_{MN+1-k_j}\}$ be the active set of items during phase k_j with decreasing true reward means i.e. $\mu_{a_1} \geq \dots \geq \mu_{a_{MN+1-k_j}}$. We assume that K' items in the top K are left to be found at the starting of phase k_j . Using the induction assumption this implies $\{a_1, \dots, a_{K'}\} \in \{1, \dots, K\}^*$ and $\{a_{K'+1}, \dots, a_{MN+1-k_j}\} \in \{K+1, \dots, MN\}$. Now there can be two types of errors.

- **Type 1 error** - An item a_l is accepted for some $l \geq K' + 1$.
- **Type 2 error** - An item a_l is rejected for some $l \leq K'$.

As done in Bubeck et al. [2013], we only show that **Type 1 error** does not occur and the other can be shown symmetrically. We know that a boundary item is accepted or rejected in phase k_j , thus a_l is a boundary item. Since a_l is not in the top K items, using Observation A.1, we get that there is some active item a_p (in the same list as a_l) also not in top K i.e. $p \geq K' + 1$ such that it has the highest empirical gap among all active items.

From here on wards our proof resembles the proof in Bubeck et al. [2013]. We can basically replace a_j in their proof with our a_p , K with MN and k with k_j , and repeat the steps that follow. However, to make it work we will have to use that a_p is not in the top K items as explained in Observation A.1. We show that $\Delta_{[MN+1-k_j]} > \max\{\mu_{a_1} - \mu_K, \mu_K - \mu_{a_{MN+1-k_j}}\}$. This cannot hold since at stage k_j since only $k_j - 1$ items have been accepted or rejected implying that $\Delta_{[MN+1-k_j]} \leq \max\{\mu_{a_1} - \mu_K, \mu_K - \mu_{a_{MN+1-k_j}}\}$. This will give us a contradiction similar to Bubeck et al. [2013]. However, to show this we need to use the implication of our observation that a_p is also not in the top K items and that it has the highest empirical mean reward i.e. $\hat{\mu}_{a_p} \geq \hat{\mu}_a$ for all $a \in A_{k_j}$. This is true since it has the highest empirical gap and it led to acceptance of a_l (see Algorithm 2).

*By abuse of notation we are using i to denote the i^{th} item from the top.

- Proof of $\Delta_{[MN+1-k_j]} \geq \mu_{a_1} - \mu_K$:

$$\begin{aligned} \hat{\mu}_{a_p, T_{k_j}} &\geq \hat{\mu}_{a_1, T_{k_j}} \\ \Rightarrow \mu_{a_p} + \frac{1}{4}\Delta_{[MN+1-k_j]} &\geq \mu_{a_1} - \frac{1}{4}\Delta_{[MN+1-k_j]} \\ \Rightarrow \Delta_{[MN+1-k_j]} &\geq \mu_{a_1} - \mu_{a_p} \end{aligned}$$

Now, since a_p is not in the top K items we know that $\mu_{a_p} \leq \mu_K \Rightarrow \Delta_{[MN+1-k_j]} \geq \mu_{a_1} - \mu_K$.

- Proof of $\Delta_{[MN+1-k_j]} > \mu_K - \mu_{a_{MN+1-k_j}}$:

Let $\sigma : \{1, \dots, MN+1-k_j\} \rightarrow A_{k_j}$ be a permutation with $\sigma(1) = p$, such that $\hat{\mu}_{\sigma(1), T_{k_j}} \geq \dots \geq \hat{\mu}_{\sigma(MN+1-k_j), T_{k_j}}$. Since a_p has the highest empirical gap in this phase, we know that,

$$\hat{\mu}_{a_p, T_{k_j}} - \hat{\mu}_{\sigma(K'+1), T_{k_j}} \geq \hat{\mu}_{\sigma(K'), T_{k_j}} - \hat{\mu}_{\sigma(MN+1-k_j), T_{k_j}} \quad (2)$$

We claim that there are at least $K' + 1$ items $(a_1, \dots, a_{K'}, a_p)$ in A_{k_j} such that their empirical mean rewards are $\geq \mu_K - \frac{1}{4}\Delta_{[MN+1-k_j]}$. This is trivially true for $a_1, \dots, a_{K'}$ since for $i \leq K'$, event η implies $\hat{\mu}_{a_i, T_{k_j}} \geq \mu_{a_i} - \frac{1}{4}\Delta_{[MN+1-k_j]} \geq \mu_K - \frac{1}{4}\Delta_{[MN+1-k_j]}$. Since a_p is not in the top K items and has empirical mean reward $\hat{\mu}_{a_p, T_{k_j}} \geq \hat{\mu}_{a_1, T_{k_j}}$, this also holds for a_p . This basically implies that both $\hat{\mu}_{\sigma(K'), T_{k_j}}, \hat{\mu}_{\sigma(K'+1), T_{k_j}}$ are $\geq \mu_K - \frac{1}{4}\Delta_{[MN+1-k_j]}$. Note that, since $\hat{\mu}_{\sigma(MN+1-k_j), T_{k_j}}$ is smallest it is $\leq \hat{\mu}_{MN+1-k_j}$ which under η , is $\leq \mu_{MN+1-k_j} + \frac{1}{4}\Delta_{[MN+1-k_j]}$. Also under η , $\hat{\mu}_{a_p, T_{k_j}} \leq \mu_{a_p} + \frac{1}{4}\Delta_{[MN+1-k_j]}$. Putting all of these together we get that

$$\begin{aligned} (\mu_{a_p} + \frac{1}{4}\Delta_{[MN+1-k_j]}) - (\mu_K - \frac{1}{4}\Delta_{[MN+1-k_j]}) &\geq (\mu_K - \frac{1}{4}\Delta_{[MN+1-k_j]}) - (\mu_{MN+1-k_j} + \frac{1}{4}\Delta_{[MN+1-k_j]}) \\ &\Rightarrow \Delta_{[MN+1-k_j]} \geq 2\mu_K - \mu_{a_p} - \mu_{MN+1-k_j} > \mu_K - \mu_{MN+1-k_j} \end{aligned}$$

where the last inequality again holds because a_p is not in the top K items i.e. $\mu_K > \mu_{a_p}$.

This completes our proof.

B DETAILED EXPERIMENTS

Regret Minimization: Here, we present simulation results for other choices of parameters M, N, K and for 100 independent trials. In Figure 1, we plot the results for synthetic bandit instance with Bernoulli rewards and with $N = 10, M = 20, K = 10$. Next, in Figure 2, we plot the results for synthetic bandit instance with Gaussian rewards and with $N = 5, M = 20, K = 10$. Finally, in Figure 3, we plot the results for semi-synthetic bandit instance with 100 clusters (i.e., 100 total number of arms) and with $N = 5, M = 20, K = 10$. Similar to those reported in the main paper, in these experiments too we observe that our algorithm *Ordered-CombUCB* fair much better than the baseline *CombUCB*.

K -Best Arm Identification: Similar to the main paper, we generate a *hard bandit instance* by sampling arm means uniformly in $[0.45, 0.55]$ and then sampling the rewards from Gaussian distributions with aforementioned means and projected to $[0, 1]$. We run our algorithm *ordered SAR* and the baseline algorithm *SAR* for rounds $T \in [1000, \dots, 10000]$. To mitigate the effect of randomness (as seen in the plots reported in the main paper), we increase number of independent trials to 1000 and plot the probability of error for both algorithms in Figure 4. Similar to those reported in the main paper, here too we find that the failure probability of *Ordered SAR* is consistently lower than that of *SAR*.

References

- Sébastian Bubeck, Tengyao Wang, and Nitin Viswanathan. Multiple identifications in multi-armed bandits. In *International Conference on Machine Learning*, pages 258–265. PMLR, 2013.
- Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvari. Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, pages 535–543. PMLR, 2015.

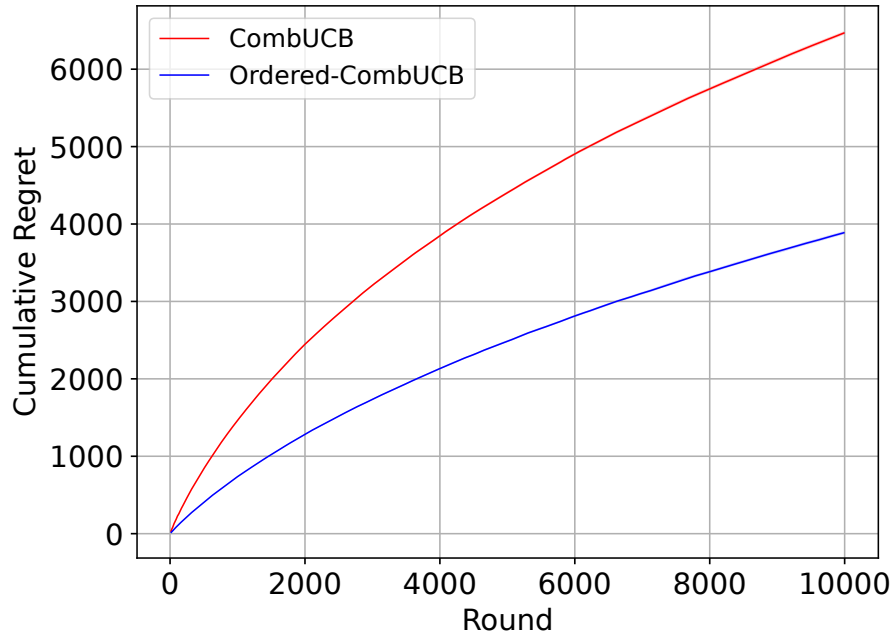


Figure 1: Comparison of cumulative regret for CombUCB and Ordered CombUCB on synthetic Bernoulli bandit instance.

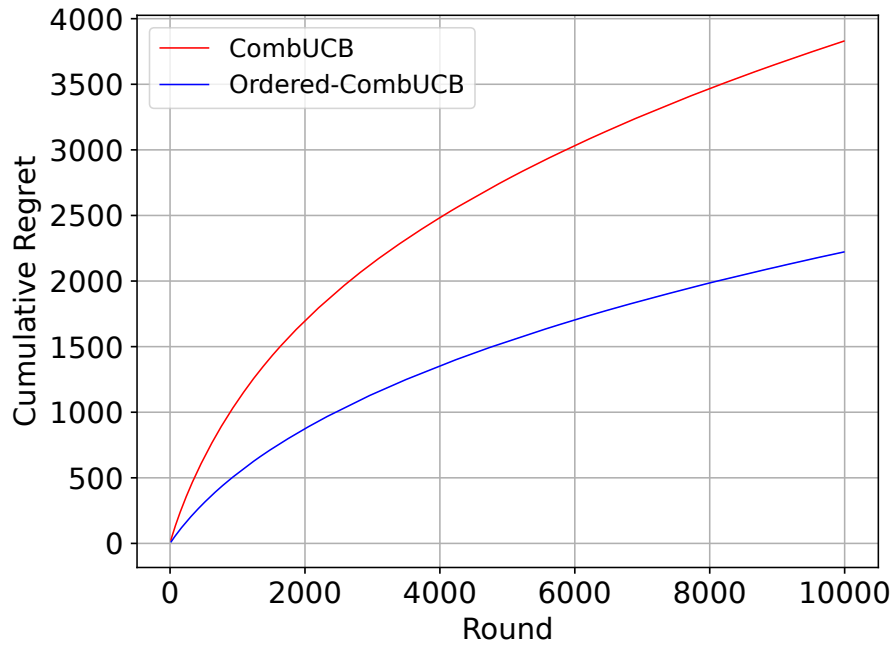


Figure 2: Comparison of cumulative regret for CombUCB and Ordered CombUCB on synthetic Gaussian bandit instance.

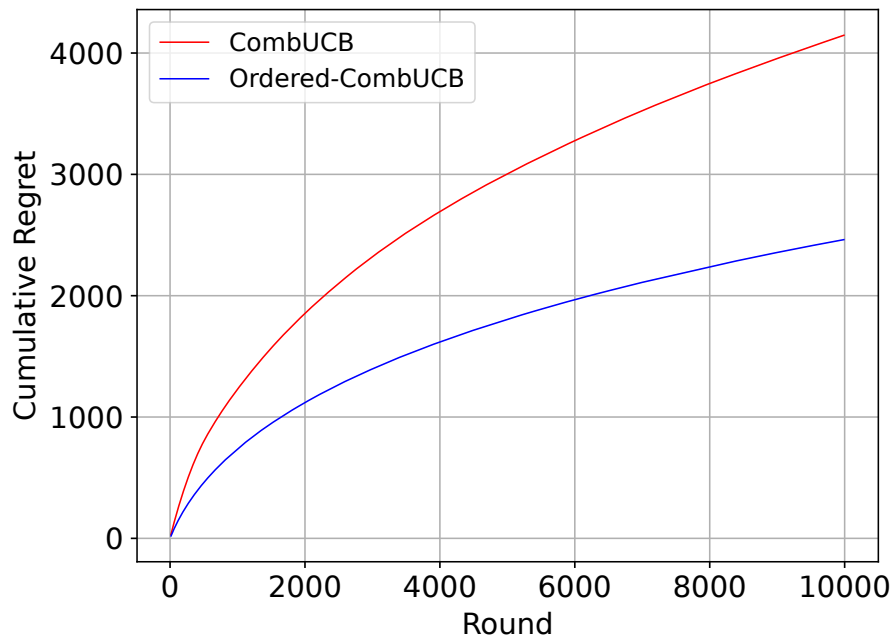


Figure 3: Comparison of cumulative regret for CombUCB and Ordered CombUCB on semi-synthetic bandit instance.

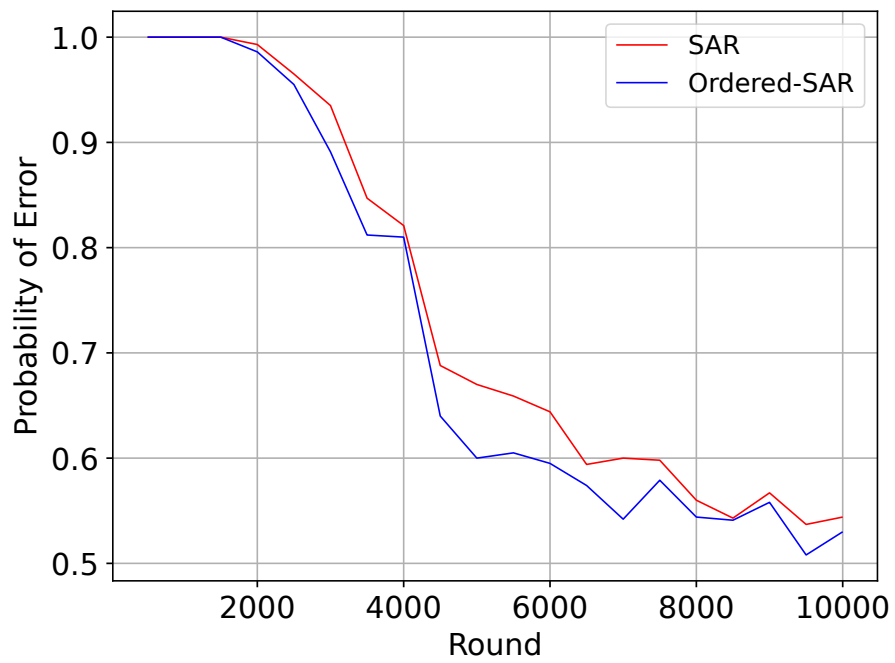


Figure 4: Comparison of probability of error for SAR and Ordered SAR on Gaussian bandit instance.