Cascading Bandits in Fixed Budget Setting

Siddharth Kalra and Shubham Patel
Prof. Subrahmanya Swamy Peruru

19th May 2025

Contents

1 Notations

- L: Number of items
- K: Number of items in a single arm
- T: Budget of arm pulls
- w_i: probability of i^{th} item being liked, items ordered such that $w_i \geq w_{i+1}$
- $\hat{w_i}$: empirical mean of i^{th} item
- $\Delta_{i,j}$: $|w_i - w_j|$
- N: Number of arm pulls taken to get one observation of each item
- round: Pulling arms until we get one more observation per item in equal-exploration algorithm
- N_i: Number of arms pulled in i^{th} round
- η: Number of rounds completed until budget exhausted
- W: $\sum_{i=1}^{N} w_i$
- $W_{i:j}$: $\sum_{k=i}^{j} w_k$
- μ_N: Upper bound on $E[N]$, equal to $W + \frac{L}{K}$
- $\mathcal{P}$: Number of phases in sequential halfing variant
- F: Event that the algorithm fails to output best-K items
- F_p: Event that in the p^{th} phase of the sequential halving variant, some of the best-K items do not survive

2 Motivation

The motivation of this problem setting can be seen to be any social media / e-commerce platform where the website recommends / shows a list different items to the user The user scrolls through them one-by-one. If the user likes a item, he clicks on its link and is redirected else he scrolls to the next item. This is repeated until either the user clicks a link or the set of related items is exhausted. The aim for the platform would be to learn and show the users those items which maximize the probability of atleast one item being clicked out of the limited size list being shown.

3 Problem Statement

There are total of L items present. Each item has a click-probability w_i, which denotes the probability that the user will click that item. An arm consists of K items $[i_1, i_2, ..., i_K]$. When an arm is pulled partial feedback from the user is received which is defined as follows. If the user clicks no item then [0,0,..,0] is the output denoting that for each item it was observed and not clicked. If the user clicks the j^{th} item in the arm, the output is [0,0,...,1,*,*,...*], which is 1 at the j^{th} denoting 1 for that j^{th} item being clicked, while 0 for items observed by user but not clicked and * for items not observed by the user. Since for the items with * we have not received any feedback about the user's behavior which is why it is a partial feedback setting. A fixed budget of T arm-pulls is provided in which the aim is to determine with high probability the best-K items with highest click-probability.

4 Related Work

Best Arm Identification for Cascading Bandits in the Fixed Confidence Setting is a paper which does analysis of cascading bandits for fixed budget setting. It proposes an elimination based algorithm in which it keeps on accepting and rejecting items wrt to the K^{th} empirical-mean item based on UCB/LCB.

Almost Optimal Exploration in Multi-Armed Bandits is a paper which proposes algorithms in normal bandit setting for fixed-confidence and fixed-budget setting. The fixed-confidence algorithm is based on medians but since it is not relevant to our setting so we don't dive much deep. The fixed-budget algorithm is Sequential Halving. The algorithm identifies the best arm by dividing budget equally into log(L) rounds. In each round it eliminates lower-half of surviving arms based on empirical estimates. It considers the samples of each round independent of the other rounds to bypass the issue of interdependence (if old samples are also considered then there might be some bias because the arms are surviving implies that they have a high estimate).

5 Equal-Exploration Algorithm

The simplest idea is to distribute budget over all items roughly equally and report the items with the top-K empirical estimates. Since this is a cascading setting so including an item in the arm does not guarantee that we would receive and observation for the item since an item before it might have been clicked by the user. So to get nearly (± 1) equal observations for all items, play an arm consisting of items which have least observations.

Algorithm:
Set estimate $\hat{w_i} = 0$ for all items
Set observation count as 0 for all items
For t from 1 to T:

- Select K the items with lowest number of observations
- Play the arm and collect output
- Update the estimates and observation counts based on the output

6 Bounding E[N]

6.1 Upper Bound on $E[N]$

The analysis of the equal-exploration algorithm would require us to get an estimate of how many observations per item do we have at the end. To determine this we would require $E[N]$ - the expected number of arm pulls it takes to get 1 observation per item. Hence here we provide a bound on it.

6.1.1 Related Stick-Breaking Problem

Problem:
Given a metallic stick which has welding joints at L points. The stick is thrown down from the top of a building. Each joint has probability of breaking as p_i. What is the expected number of pieces of the stick?

Solution:
Define X_i as the random variable which is 1 if i^{th} joint breaks and 0 if i^{th} joint does not break. Observe carefully $X = 1 + \sum_i X_i$ is the random variable denoting the number of pieces of stick. Using linearity of expectation $E[X] = \sum_i p_i$.

6.1.2 Extending the solution to cascading bandits

Define X_i as the random variable which is 1 if i^{th} item is clicked by the user and 0 if i^{th} if it is not clicked. Here $\sum_i X_i$ would not be exactly equal to the number of arms required this is because there is an additional constrain here.

The constrain is that if we have K consecutive 0's then also a new arm would be required since that maximum items in an arm is K, but this would not be counted in $\sum_i X_i$. To account for this, the term of $\frac{L}{K}$ is added which takes into account that at every K items, atleast one arm is added. Hence $\mu_N = \sum_i X_i + \frac{L}{K}$ would be an upper on N. Hence

$$E[N] \leq \frac{L}{K} + \sum_i w_i = \mu_N \tag{1}$$

6.2 Tightness of upper bound

While we have got an upper bound on $E[N]$, but if the bound is very lose it would not be much beneficial. Here we prove the tightness of the calculated upper bound and show it to lie within a factor of 2.

$\frac{L}{K} \leq N$ since atleast these many arms must be played.
From which the following can be a lower bound on $E[N]$

$$\frac{L}{K} \leq E[N] \tag{2}$$

Another lower bound on $E[N]$ would be the following

$$\sum_i w_i \leq E[N] \tag{3}$$

since LHS assumes that arm can have size greater than K but that is not the case so more arms would be needed.

Using Eqn 1. and Eqn 2. and defination of μ_N

$$\mu_N = \frac{L}{K} + \sum_i w_i \leq 2E[N] \tag{4}$$

Hence the upper bound of μ_N on $E[N]$ can be shown to be tight within a factor of 2 as

$$E[N] \leq \mu_N \leq 2E[N] \tag{5}$$

7 Analysis For Equal-Exploration Algorithm

Our algorithm is to play for as much rounds as possible in given budget. Let N_i be the number of arm pulls required to complete one round.

7.1 Good Event

Given η, the number of times each item is observed, we would then calculate the probability of failure. I.e. when our algorithm gives final output of a non-optimal arm instead of reporting the arm with best-K items.

For this, initially the criteria we had used was that for every item i in the best-K items and every item j in non-optimal items $\hat{w}_i \geq \hat{w}_j$. This required $K*(L-K)$ conditions. We then came up with a weaker criteria.

If $\hat{w}_i \geq \hat{w_K}$ for all optimal items and $\hat{w_K} \geq \hat{w}_j$ for all non-optimal items, then also it is sufficient. This reduces the required coniditions from $K*(L-K)$ to only L.

A necessary and sufficient criteria could be that $\min \hat{w}_i \geq \max \hat{w}_j$ but the expressions could be more complicated than the previous one. Hence we stick with the previous one.

Using the inequality from Quiz 1 (sub-gaussian r.v.):

$$P(\hat{w}_i < \hat{w}_j|\eta) \leq e^{-\frac{\eta\Delta_{i,j}^2}{4}}$$

$$P(F|\eta) \leq \sum_{i \neq K} e^{-\frac{\eta\Delta_{i,K}^2}{4}} \tag{6}$$

7.2 McDiarmid Inequality

Consider $L \times \eta$ number of bernoulli random variables, giving the outcome of each of the L items for η rounds. Now, $N_1 + N_2 + ... + N_\eta$, i.e. total number of arm pulls required, is a function of these $L \times \eta$ random variables. Below we show that changing any of these random variables will change the output of $\sum_{i=1}^{N} N_i$ by at max 1.

$$P\left(\sum_{i=1}^{\eta} N_i - E\left[\sum_{i=1}^{\eta} N_i\right] \geq \epsilon\right) \leq e^{\frac{-2\epsilon^2}{L\eta}} \tag{7}$$

Examples

For examples let K = 3 and L = 5
Let $\hat{W} = (\hat{w}_1, \hat{w}_2, ..., \hat{w}_L)$ denote the observations, 1 per each item.

$N = f(\hat{w_1}, \hat{w_2}, ..., \hat{w_L})$
Changing $\hat{w_i}$ causes N to change by atmost 1.

Change $\hat{w_i}$ from 0 to 1:
If i is the last item of an arm then no change in N

- E.g. $i = 3$,
  $\hat{W}_{before} = (0, 0, \mathbf{0}, 0, 0), N_{before} = 2$,
  $\hat{W}_{after} = (0, 0, \mathbf{1}, 0, 0), N_{after} = 2$

Else $+1$ in N or no change in N.

- E.g.1. $i = 1$,
  $\hat{W}_{before} = (\mathbf{0}, 0, 0, 0, 0), N_{before} = 2$,
  $\hat{W}_{after} = (\mathbf{1}, 0, 0, 0, 0), N_{after} = 3$
- E.g.2. $i = 2$,
  $\hat{W}_{before} = (0, \mathbf{0}, 0, 0, 0), N_{before} = 2$,
  $\hat{W}_{after} = (0, \mathbf{1}, 0, 0, 0), N_{after} = 2$

Change $\hat{w_i}$ from 1 to 0:
If i is the last item of an arm then no change in N.

- E.g. $i = 3$,
  $\hat{W}_{before} = (0, 0, \mathbf{1}, 0, 0), N_{before} = 2$,
  $\hat{W}_{after} = (0, 0, \mathbf{0}, 0, 0), N_{after} = 2$

Else -1 in N or no change in N.

- E.g.1. $i = 1$,
  $\hat{W}_{before} = (\mathbf{1}, 0, 0, 0, 0), N_{before} = 3$,
  $\hat{W}_{after} = (\mathbf{0}, 0, 0, 0, 0), N_{after} = 2$
- E.g.2. $i = 2$,
  $\hat{W}_{before} = (0, \mathbf{1}, 0, 0, 0), N_{before} = 2$,
  $\hat{W}_{after} = (0, \mathbf{0}, 0, 0, 0), N_{after} = 2$

7.3 Computing $P(\eta < \eta_0)$

Let us forget the budget for once, and consider that we play for η_0 rounds. We try to get an upper bound on cumulative distribution of η_0.

Consider $L \times \eta_0$ number of bernoulli random variables, giving the outcome of each of the L arm for η_0 rounds. Now, $N_1 + N_2 + ... + N_{\eta_0}$, i.e. total number of arm pulls required, is a function of these $L \times \eta_0$ random variables. Also we had previously shown that changing any of these random variables will change the output of $\sum_{i=1}^{N} N_i$ by at max 1. Hence, using McDiarmid concentration inequality, we have,

$$P\left(\sum_{i=1}^{\eta_0} N_i - E\left[\sum_{i=1}^{\eta_0} N_i\right] \geq \epsilon\right) \leq e^{\frac{-2\epsilon^2}{L\eta_0}}$$

$$\implies P\left(\sum_{i=1}^{\eta_0} N_i - \left(\eta_0 W + \eta_0 \frac{L}{K}\right) \geq \epsilon\right) \leq e^{\frac{-2\epsilon^2}{L\eta_0}}$$

using the fact that $E[N_i] \leq W + \frac{L}{K}$, where $W = \sum_{i=1}^{N} w_i$.

Now, to take budget into account, we can choose $\epsilon = T - (\eta_0 W + \eta_0 \frac{L}{K})$ since this is the largest ϵ that will ensure that $\sum_{i=1}^{\eta_0} N_i$ remains less than T given the event $\sum_{i=1}^{\eta_0} N_i - \left(\eta_0 W + \eta_0 \frac{L}{K}\right) \geq \epsilon$ does not occur.

Using this, we obtain, $P\left(\sum_{i=1}^{\eta_0} N_i \geq T\right) \leq e^{\frac{-a+bT-cT^2}{L\eta_0}}$ for some constants a, b, c. Note that the term $P\left(\sum_{i=1}^{\eta_0} N_i \geq T\right)$ is equal to $P(\eta < \eta_0)$

7.4 Considering η_0 as analysis parameter

The idea is similar to the derivation chernoff bound. In that λ is a parameter of the analysis which is then chosen such that the upper bound is tightest. Similarly here we consider η_0 as a parameter of the analysis and not something the algorithm requires to be performed.

Define event F: The algorithm fails to output best-K items

$$P(F) = \sum_{i=1}^{\eta_{\max}} P(F|\eta = i)P(\eta = i)$$

For a fixed constant η_0, we can write,

$$\sum_{i=1}^{\eta_{\max}} P(F|\eta = i)P(\eta = i) = \sum_{i=1}^{\eta_0 - 1} P(F|\eta = i)P(\eta = i) + \sum_{i=\eta_0}^{\eta_{\max}} P(F|\eta = i)P(\eta = i)$$

$$\leq \sum_{i=1}^{\eta_0 - 1} P(\eta = i) + \sum_{i=\eta_0}^{\eta_{\max}} P(F|\eta = i)P(\eta = i)$$

$$= P(\eta < \eta_0) + \sum_{i=\eta_0}^{\eta_{\max}} P(F|\eta = i)P(\eta = i)$$

$$\leq e^{\frac{-a+bT-cT^2}{L\eta_0}} + \sum_{i=\eta_0}^{\eta_{\max}} P(F|\eta = i)P(\eta = i)$$

$$\leq e^{\frac{-a+bT-cT^2}{L\eta_0}} + \sum_{i=\eta_0}^{\eta_{\max}} P(F|\eta = \eta_0)P(\eta = i)$$

$$\leq e^{\frac{-a+bT-cT^2}{L\eta_0}} + P(F|\eta = \eta_0) \sum_{i=\eta_0}^{\eta_{\max}} P(\eta = i)$$

$$\leq e^{\frac{-a+bT-cT^2}{L\eta_0}} + P(F|\eta = \eta_0)$$

Now, from the fact that we can use Hoeffding inequality for any pair of the items, even if it is conditioned on event $\eta = \eta_0$ we can write the above expression as,

$$P(\text{wrong answer}) \leq e^{\frac{-a+bT-cT^2}{L\eta_0}} + \sum_{i=1, i\neq K}^{L} e^{-\frac{1}{4}\eta_0 \Delta_{iK}^2}$$

We can optimize this with respect to η_0 to obtain the best η^* to set for our algorithm and obtain an upper bound for probability of failure.

8 Inter-dependence

8.1 Challenge

There is an important aspected which had been not discussed in detail regarding the analysis done till now. The $P(F|\eta)$ has been used by us in calculating the failure probability of the algorithm. N, and hence also η, is a function of the observation of items, since more 0s observed would imply smaller N and vice-versa. A complication here arises because now we are conditioned on the event that N, hence also η, is within a given range. This implies that there is a certain dependence now on the $\hat{w_i}$. So the analysis for calculating $P(\hat{w_i} > \hat{w_j})$ which was done initially might not be directly applicable since it assumed the samples to be independent but here they are conditioned on the event of value of N, η.

8.2 Why analysis is valid when conditioned on $\eta = i$

As mentioned in Section 1.3.1 and Lemma 6.2, we can use Azuma Hoeffding inequality to apply concentration inequality to a single item, even for the case when the number of samples obtained is a function of the history. Though we are not sure whether this holds even when the deviation probability is not kept constant (independent of number of samples) but based on the above we feel that it could hold. This is yet to be shown rigorously.

9 Sequential Halving for Cascading Bandits

Based on the algorithm of the second paper, we adapt the sequential halving algorithm for cascading bandits.

9.1 Algorithm

Total phases: $\mathcal{P} = \log_2(L - K)$
Surviving items: $S = [L]$
For each phase p from 1 to $\mathcal{P}$:

- Set $\hat{w_j}, n_t(j) = 0 \; \forall \; j \;\; \in S$
- for t from 1 to T:
  - Play an arm of K items with smallest n_t
  - Update $\hat{w_j}$ and $n_t(j)$
- Remove $\frac{L-K}{2^{p+1}}$ arms with smallest $\hat{w_j}$ from S

Report S

9.2 Analysis

Define Events:

- F - Algorithm fails to output best-K items
- F_p - Some best-K items do not survive in phase p

$$P(F) = \cup_p P(F_p)$$

Each phase of the above algorithm can be viewed as one instance the equal-exploration algorithm.

9.2.1 Single Phase

Consider the p^{th} phase.
The worst case for the algorithm would be when the easiest items are eliminated first and the harder items remain.
Define index $a_p^* = L - \frac{L-K}{2^{p+1}}$. This would be the item having largest w_i out of the items we want to eliminate.
It would be sufficient to ensure that $\hat{w_i} \geq \hat{w_{a_p^*}}$ for better items and $\hat{w_j} \leq \hat{w_{a_p^*}}$ for the other items being eliminated.

$$P(F_p) \leq \sum_{i=1}^{a_p^*-1} P(\hat{w_i} < \hat{w_{a_p^*}}) + \sum_{j=a_p^*+1}^{a_{p-1}^*} P(\hat{w_j} > \hat{w_{a_p^*}})$$

Using argument similar as equal-exploration for good events and considering η_p as a parameter of analysis and using independence

$$P(F_p) \leq e^{-\frac{(T_p - \eta_p \mu_p)^2}{2 L_p \eta_p}} + \sum_{i \neq a_p^*}^{a_{p-1}^*} e^{-\frac{\eta_p}{2} \Delta_{i,a_p^*}^2}$$

where $T_p = \frac{T}{\mathcal{P}}, L_p = L - \frac{L-K}{2^{p+1}}, \mu_p = \frac{L_p}{K} + W_{1:L_p}$

9.2.2 All Phases

$$P(F) \leq \sum_p P(F_p)$$

$$P(F) \leq \sum_p (e^{-\frac{(T_p - \eta_p \mu_{\mathcal{P}})^2}{2L_p \eta_p}} + \sum_{i \neq a_p^*}^{a_{p-1}^*} e^{-\frac{\eta_p}{2}\Delta^2_{i,a_p^*}})$$

The above expression holds for all valid η_p. By performing an optimization over these one can get the corresponding η_p^* denoting the tightest analysis.

10 Comparison with Multiarm Bandits

We evaluate the results at $\eta = \frac{T}{2\mu_N}$ (approx half of E[η] bound). The motivation for this comes from the fact that $E[\eta] > \frac{T}{\mu_N} - 1$ which is an inequality that comes from the domain of stopping time analysis and which we have used as it is without diving into the derivation.

10.1 Equal-Exploration Algorithm

For $K = 1$, we obtain $\frac{T}{L}$ observations of each item. In this multi-arm setting, we can separate the items with probability of $\sum_{i \neq K}^{L} e^{-\frac{T\Delta^2_{iK}}{4L}}$. This is approximately of the order of the second term outputted by the cascading analysis - $Le^{-\frac{T\Delta^2}{4(L+W)}}$. They are of same order because $W \leq L$. The first term outputted by our analysis ($P(\eta < \frac{T}{2\mu_N}) \leq e^{-\frac{T(W+\frac{L}{K})}{L}}$) is smaller as compared to these terms hence can be ignored. Hence both the actual and output by algorithm are of similar order.

Algorithm Output $\approx$ Actual:

$$e^{-\frac{T(W+\frac{L}{K})}{L}} + \sum_{i \neq K}^{L} e^{-\frac{T\Delta^2_{iK}}{4(L+W)}} \approx \sum_{i \neq K}^{L} e^{-\frac{T\Delta^2_{iK}}{4L}}$$

10.2 Sequential Halving

For K = 1, similarly the term of $P(\eta < \eta_p)$ would be much smaller as compared to the probability of $\hat{w}_i < \hat{w}_j$ at $\eta = \frac{T_p}{2\mu_p}$. Hence simply following the analysis of the paper a very similar analogy as equal-exploration algorithm will hold here.

11 Future Directions

Beyond this project

11.1 Theoretical Lower Bounds

Currently, algorithms have been proposed for both fixed-budget and fixed-confidence settings in this work and in previous works. Also, the upper bounds of their performance have been presented. But one thing that could be further explored upon is deriving lower-bound for the problem setting. Just like in normal bandits, there have been bounds derived for how bad the best algorithm would perform; a similar direction can be explored in the domain of cascading bandits, similar to those discussed at the start of this paper for normal bandits.

This would help one better assess how well the proposed algorithms are performing as compared to the theoretically best algorithm. It would also help one better understand the intrinsic difficulty of these problems. In the normal case (where we try to minimize regret), there have been works in the cascading setting like [paper] derive lower bounds. But in the fixed-confidence/budget setting, this is a requiring more work.

11.2 Approximation Algorithms

The algorithm we have aimed to create in the fixed budget setting is to give the "best-K" items from the whole set. If we look at the literature on non-cascading bandits, there are variants that allow the ϵ threshold of error. For example, the median algorithm related is presented in the second paper of related work. Also, in a cascading setting, the Vincent tan paper begins with an ϵ variant. So, similar extensions would be an interesting direction to explore in the fixed-budget setting since, in real-world scenarios, there is a certain tolerance for error, which would be fine.

11.3 Structured Bandits: Non-independent arms

In the current scenario, we have taken that for different arms, the outcomes are independent, but this might not be the case for situations where the outcomes have influence over each other. In normal bandits, there have been works like [paper] analyzing the setting of fixed-budget for structural bandits. The case tackled in this work for cascading bandits was on unstructured bandits, while the more general case of structured bandits seems to not have been explored yet for cascading settings and could be a future direction that can be explored.

11.4 Incorporating states - Contextual Bandits - RL

In the discussed setting, the focus was on multi-arm bandits with only one state; while this is a nice model for a single user, it is not scalable for practical settings like social media platforms where there are multiple users. So, to make the problem setting more practical it would be interesting to study the fixed-confidence/budget in contextual bandits. There are some past works studying cascading + contextual [paper] but seem to have not focused on both fixed-confidence/budget settings. Another extension of this would be to the full RL setting, where the state space is not discrete and is continuous. Similarly, there are some past works studying cascading RL [paper] but seem to have not focused on both fixed-confidence/budget settings. So this could also be a direction on which future work can be built.

11.5 Hybrid settings

Some previous works focus on mixing settings like federated learning with cascading bandits like [paper]. Also, currently, the cascading bandits treat all positions as the same, i.e., ideally, a user might get bored by scrolling, so they would be more likely to click an item appearing at the top of the list rather than at the end. Such factors can be taken into account, making the problem more realistic and helping in more accurately predicting trends observed and improving recommendation systems.

11.6 Analysis of another improved variant

Similar to the algorithm presented in the first paper for fixed confidence setting, we can develop an improved variant based on elimination. This could be another algorithm that can be analyzed.

Description:
In a given round we play arms similar to before until we get one observation per item. For each item we maintain a probability estimate, LCB and UCB. At the end of each round we reject those arms which are most-likely non-optimal and accept those arms which are most-likely optimal. This decision is taken based on the k^{th} largest estimate's UCB and LCB. This process is repeated until time budget is exhausted or only K items remain. Finally the items with top-K estimates are reported.

Algorithm:
k = K
Repeat until K-items Accepted or L-K items rejected or time-budget exhausted

1. Get one observation per item using N arms for the surviving items
2. Update the $\hat{w}_i, UCB_i, LCB_i$ with the new observation
3. $i^* = k^{th}$ biggest $\hat{w}_i$
4. Reject all i such that $UCB_i < LCB_{i^*}$
5. Accept all i such that $LCB_i \geq UCB_{i^*}$

6. Update k by subtracting number of items accepted in this round
7. Update surviving items by removing the rejected and accepted

If accepted items less than K, add top-k $\hat{w_i}$ items to accepted Report the Accepted items

Within this project

11.7 Making exponent tighter from $\frac{1}{4}$

To find $P(\hat{w_i} < \hat{w_j})$, where $w_i > w_j$ and number of observations $= \eta$.
Currently we use $P(\hat{w_i} < \hat{w_j}) \leq e^{-\frac{1}{4}\eta\Delta_{i,j}^2}$. But it might be possible that the factor of $\frac{1}{4}$ could be reduced to $\frac{1}{2}$ or 1 (as has been used in the analysis for sequential halving by the paper).

11.8 Rigorously showing Independence

As discussed previously there was the issue of independence of samples which could impact our analysis. For this we went through the two mentioned resources. Based on them, it seemed that if they were true, then even in our slightly different case, independence would hold, but this is yet to be shown formally for this setting.

11.9 Understanding $E[\eta] > \frac{T}{\mu_N} - 1$

For the purposes of the project, we took this inequality as 'given'. To get a deeper understanding of the problem on can dive further into the proof of this inequality.