Improved Regret Bounds with Clustered UCB

Abstract

In this paper, we present a novel algorithm which achieves a better upper bound on regret for the stochastic multi-armed bandit problem then some of the existing algorithms. Our proposed method clusters arms based on their average estimated payoff. This results in a more efficient elimination of arms within clusters as well as the deletion of weak clusters. We prove the regret upper bound as $\sum_{i \in B_m} \left(\max \left\{ \left(\frac{32}{(\Delta_i)^3} \right), \left(\frac{25\Delta_i}{(\Delta^2)(0.16T\Delta^2)^{2|B_m|^2\Delta/5}} \right) \right\} + \left(\Delta_i + \frac{32 \log (T \frac{\Delta_i^4}{16})}{\Delta_i} \right) \right)$, Δ is the minimal gap between the means of the reward distributions of the optimal arm and a sub-optimal arm, A is the set of arms, T is the horizon and $\psi(m)$ is a parameter of the problem. This bound improves upon the existing algorithm of UCB-Revisited, MOSS, KL-UCB and UCB1 under certain cases. We corroborate our findings both theoretically and empirically and show that by sufficient tuning of parameter, we can achieve a lower regret than all the algorithms mentioned. In particular, in the test-cases where the arm set is dominated by small Δ and large K we achieve a significantly lower cumulative regret than these algorithms.

1 Introduction

In this paper, we address the stochastic multi-armed bandit problem, a classical problem in sequential decision making. Here, a learning agent is provided with a set of choices or actions, also called arms and it has to choose one arm in each timestep. After choosing an arm the agent receives a reward from the environment, which is an independent draw from a stationary distribution specific to the arm selected. The goal of the agent is to maximize the cumulative reward. The agent is oblivious to the mean of the distributions associated with each arm, denoted by r_i, including the optimal arm which will give it the best average reward, denoted by r^*. The agent records the cumulative reward it has collected for any arm divided by the number of pulls of that arm which is called the estimated mean reward of an arm denoted by $\hat{r}_i$. In each trial the agent faces the exploration-exploitation dilemma, that is, should the agent select the arm which has the highest observed mean reward till now (exploitation), or should the agent explore other arms to gain more knowledge of the true mean reward of the arms and thereby avert a sub-optimal greedy decision (exploration). The objective in the bandit problem is to maximize cumulative reward which will lead to minimizing the expected cumulative regret. We define the expected regret (R_T) of an algorithm after T trials as

$$\mathbb{E}[R_T] = r^*T - \sum_{i \in A} r_i \mathbb{E}[N_i]$$

where N_i denotes the number of times the learning agent chooses arm i within the first T trials. We also define $\Delta_i = r^* - r_i$, that is the difference between the mean of the optimal arm and the i-th sub-optimal arm (for simplicity we assume that there is only one optimal arm which will give the highest payoff). The problem, gets more difficult when the Δ_i's are smaller and arm set is larger. Also let $\Delta = min_{i \in A} \Delta_i$, that is it is the minimum possible gap over all arms in A.

Related work

This section is too long and lacks a qualitative comparison with Clus-UCB. If the target is one of the ML conferences, then this section has to restricted to a couple of paragraphs that compare Clus-UCB with closely related works, for e.g. UCB-revisited.

Bandit problems have been extensively studied under various conditions. One of the first works can be seen in Thompson (1933), which deals with choosing between two treatments to administer on patients who come in sequentially. Further studies by Robbins (1952) and Lai and Robbins (1985), established an asymptotic lower bound for the regret. Lai and Robbins proved in Lai and Robbins (1985) that for any allocation strategy and for any sub-optimal arm i, the regret is lower bounded by

$$\lim_{T\to\infty} \inf \frac{\mathbb{E}[R_T]}{\log T} \geq \sum_{i:r_i<r^*} \frac{(r^* - r_i)}{D(p_i||p^*)},$$

where $D(p_i||p^*)$ is the Kullback-Leibler divergence over the reward density p_i and p^* over the arms having mean r_i and r^*.

Of the several algorithms mentioned in Auer et al. (2002a), UCB1 has a regret upper bound of $O\left(\frac{K \log T}{\Delta}\right)$ whereas in the same paper the author's propose UCB2 algorithm which has a tighter regret bound than UCB1. UCB2 has a parameter α that needs to be tuned and its regret is upper bounded by $\sum_{i\in A}\left(\frac{(1+C_1(\alpha))\log T}{2\Delta_i}\right) + C_2(\alpha)$, where $C_1(\alpha) > 0$ is a constant that can get arbitrarily small when α gets smaller but $C_2(\alpha)$ consequently starts increasing. Another type of strategy was first proposed by Sutton and Barto (1998) called $\epsilon-$greedy strategy where the agent behaves greedily most of the time pulling the arm having highest $\hat{r}_i, \forall i \in A$ and sometimes with a small probability ϵ it will try to explore by pulling a sub-optimal arm. In Auer et al. (2002a) they further refined the same algorithm and proposed ϵ_n-greedy with regret guarantee. For the ϵ_n-greedy it is proved that if the parameter ϵ, is made a function over time, like $\epsilon_t = \frac{const.K}{d^2 t}$, such that $0 < \epsilon < 1$, then the regret grows logarithmically $\left(\frac{K \log T}{d^2}\right)$. This algorithm performs well given that $0 < d < \min_{i\in A} \Delta_i$ and for large $const$ value the result actually becomes stronger than UCB1. For further insight into various approaches in dealing with stochastic multi-armed bandit we refer the reader to Bubeck and Cesa-Bianchi (2012). In Auer and Ortner (2010) they prove an improved regret bounds for the algorithm UCB-Revisited in the order of $\sum_{i\in A}\left(const * \frac{K \log(T\Delta_i^2)}{\Delta_i}\right)$ for very small Δ_i over a larger set of arms. This is a round based method, where in every round, all the arms are pulled an equal number of times, then based on certain conditions they eliminate some arms and this goes on till one arm is left. In this paper we refer to these type of algorithms as round based algorithms.

Other prominent round based elimination algorithms include Successive Reject, Successive Elimination and Median Elimination. The Successive reject algorithm was proposed by Audibert and Bubeck (2010), where they explore the pure exploration scenario in a fixed budget/horizon setup. Successive Reject tries to proceed in a phase-wise manner eliminating one arm after each phase. They try to bound the regret by defining two hardness parameter $H_1 = \sum_{i=1}^{K} \frac{1}{\Delta_i^2}$ and $H_2 = \max_{i\in A} i\Delta_i^{-2}$. However, knowledge of these two parameters beforehand is a difficult task so an online approach to estimate them is used in Adaptive UCB-E. The next algorithm called Successive Elimination proposed by Even-Dar et al. (2006) also proceeds by eliminating one arm after every round and the authors give PAC-guarantees for them. In PAC guarantee algorithms the learning agent comes up with an ϵ-optimal arm with δ error probability. They also propose a further modification of Successive Elimination called Median Elimination which removes one half of the arms after every round. The sample complexity of Successive Elimination for any sub-optimal arm is bounded by $O\left(\sum_{\Delta_i>\epsilon} \frac{\log(\frac{K}{\delta\Delta_i})}{\Delta_i^2} + \frac{N(\Delta,\epsilon)}{\epsilon^2}\log(\frac{N(\Delta,\epsilon)}{\delta})\right)$, where $N(\Delta,\epsilon)$ are the number of arms which are $\epsilon-$optimal whereas for Median Elimination the sample complexity for any sub-optimal arm is bounded

by $O(\sum_{i=1}^{\log_2 K} \frac{K}{\epsilon^2} \log(\frac{1}{\delta}))$, where δ and ϵ are the parameters defined before. For further insight into various approaches in dealing with stochastic multi-armed bandit we refer the reader to Bubeck and Cesa-Bianchi (2012).

It is also important to distinguish between two approaches in the UCB type algorithms. One approach uses explicit mean estimation using Chernoff Bounds for calculating the confidence bound whereas the other approach uses variance estimation using Bernstein or Bennett's Inequality to estimate the confidence bound. Such variance estimation is found in Auer et al. (2002a) for the UCB-Normal and UCB-Tuned algorithm and later in Audibert et al. (2009) where they use Bernstein Inequality to build the confidence term for UCB-V algorithm. But UCB1, UCB2, MOSS, UCB-Revisited, KL-UCB and Median-Elimination uses no such variance estimation techniques. In our analysis also we use no variance estimation methods.

Some of the more recent algorithm like MOSS and KL-UCB provides further refinements to the upper bound in the stochastic multi-armed bandit case. In Audibert and Bubeck (2009) the MOSS(Minimax Optimal Strategy in Stochastic case) algorithm achieves a distribution free upper bound on the regret as $const.\sqrt{TK}$ and the authors also propose a distribution dependent upper bound as $\sum_{i:\Delta_i>0} \frac{K \log(2 + T\Delta_i^2/K)}{\Delta_i}$. In Garivier and Cappé (2011) the authors propose KL-UCB which achieves an upper bound on regret as $\sum_{i:\Delta_i>0} \left(\frac{\Delta_i(1+\alpha)\log T}{D(r_i, r^*)} + C_1 \log\log T + \frac{C_2(\alpha)}{T^{\beta(\alpha)}} \right)$ which is strictly better than UCB1 as we know from Pinsker's inequality $D(r_i, r^*) > 2\Delta_i^2$. KL-UCB beats UCB1, MOSS and UCB-Tuned in various scenarios. Another algorithm that has been proposed in Abbasi-Yadkori et al. (2011) called $UCB(\delta)$ creates a confidence interval which does not depend on timestep or on horizon T. The regret upper bound in this algorithm is given by $\sum_{i:\Delta_i>0} \left(3\Delta_i + \frac{16}{\Delta_i} \log\left(\frac{2K}{\Delta_i \delta}\right) \right)$, where δ is the error probability as defined before. In Honda and Takemura (2010) the authors come up with the algorithm Deterministic Minimum Empirical Divergence also called DMED+(as referred by Garivier and Cappé (2011)) which is first order optimal. This algorithm keeps a track of arms whose empirical mean are close to the optimal arm and takes help of large deviation ideas to find the optimal arm.

Also, we mention the algorithm Exp3 in the adversarial bandit scenario. In the adversarial case, the agent is playing against an adversary who can arbitrarily set a reward on any arm. Algorithms used in adversarial case can be used in the stochastic scenario but not vice-versa. For Exp3 in Auer et al. (2002b) the authors achieve an upper bound of the order of $O(S\sqrt{KT \log KT})$, where S is defined as the hardness of a problem. Finally, we mention one algorithm that involves Bayesian estimation technique called Thompson Sampling(TS) which uses a prior distribution over arms to estimate the posterior distribution and thereby converge on the optimal arm. In Agrawal and Goyal (2011) the authors come up with a regret upper bound of the order of $O([K(\frac{1}{\Delta^2})^2] \log T)$ for K-arm stochastic bandits using Thompson Sampling.

Our contributions

In this work, we present a novel method where we cluster the arms in each round based on the average estimated payoff $\hat{r}_i$. The purpose of this approach is to control exploration and at the same time exploit the clusters formed to create tight confidence bounds. To do this, we use hierarchical agglomerative clustering using the single-linkage clustering scheme as mentioned in Friedman et al. (2001). This is used to group the arms together at the start of each round. Then we deploy a two-pronged approach of exploring inside each clusters to eliminate sub-optimal arms and separately based on other conditions eliminating some weak clusters with all the arms inside it. After each round, at the beginning of the next round we again cluster arms after destroying the old cluster structures formed in the previous round. The logic behind clustering at the beginning of each round afresh is simply that at the initial rounds we have clusters formed with very

bad purity level(we can imagine the purity level of a cluster being judged by how many arms having similar or ϵ_m-close means r_i getting clustered together), where in the later rounds we can have tight clusters with high purity level since now we have a better estimate of $\hat{r}_i, \forall i \in A$.

The within cluster arm elimination condition and the entire cluster elimination are two complimentary strategies for speedy elimination of sub-optimal arms. Unlike UCB-Revisited from Auer and Ortner (2010) the within cluster arm elimination leads to dividing the larger problem of finding the optimal arm from the whole arm set into small sub-problems where in each such small clusters (say M clusters, $M \leq K$) we have M arm elimination conditions, thereby increasing the probability of deletion of a sub-optimal arm and hence reducing regret. Also by dividing it into small sub-problems the growth of our pulls in each round is always small given the optimal arm has survived which we ensure by tuning parameters appropriately. We start with a very small cluster size limit(say 2) and double the limit after every round. This cluster size limit always remains bounded since we are exploring each arm in a cluster based on its size, such bounds helps in controlling the exploration within a cluster because we know that single link clustering often forms large chains where the first and large elements may not at all be similar to each other. The entire cluster elimination conditions exploit the idea that if the optimal arm has survived till the later rounds it will be in a cluster of its own with no other sub-optimal arm and then we eliminate all sub-optimal clusters in a single round. This is essentially the same core concept of various round based strategies.

Summarizing, the contributions of this research are listed below:

1. We propose a cluster based round-wise algorithm with two arm elimination conditions in each round.

2. We achieved a lower regret bound than UCB-Revisited(Auer and Ortner (2010)), UCB(δ) (Abbasi-Yadkori et al. (2011)), UCB1(Auer et al. (2002a)), UCB2(Auer et al. (2002a)), UCB-Tuned(Auer et al. (2002a)), Median Elimination(Even-Dar et al. (2006)), Exp3(Auer et al. (2002b)) and MOSS (Audibert and Bubeck (2009) in scenarios when Δ is small and K is large which is encountered frequently in web-advertising domain (as stated in Garivier and Cappé (2011)), which we verify empirically.

   I don't see a rigourous theoretical justification for this claim anywhere in the paper. We have a regret bound in Theorem 1, but where is it shown that the bound in Thm 1 is better than UCB-Revisited(Auer and Ortner (2010)), UCB(δ) (Abbasi-Yadkori et al. (2011)), UCB1(Auer et al. (2002a)), UCB2(Auer et al. (2002a)), UCB-Tuned(Auer et al. (2002a)), Median Elimination(Even-Dar et al. (2006)), Exp3(Auer et al. (2002b)) and MOSS (Audibert and Bubeck (2009)?

   "when Δ is small and K is large which is encountered frequently in web-advertising domain..." Can you give a reference that justifies low Δ in web-advertising? (Subho) Reference given, rest of the contribution unchanged, to be decide after the regret bound is found

3. Our algorithm also empirically compares well with DMED, DMED(+) and KL-UCB.

   Is the comparison empirical or theoretical? Please specify . Done. written empirically, Subho

4. In the critical case when $r_1 = r_2 = .. = r_{K-1} < r^*$ and Δ_is are small, this approach has a significant advantage over other methods.

5. We also come up with an error bound to prove that the error probability decreases exponentially after each round.

6. Unlike KL-UCB our algorithm parameter $\psi(m)$ is not distribution-specific and also our algorithm does not involve calculation of a complex, time consuming function like the divergence function of KL-UCB.

The paper is organized as follows, in section 2 we present notations and preliminary assumptions. In section 3 we present the algorithm and discuss why it works. Section 4 deals with all the proof including the proofs on regret bound. In section 5 we present the experimental run of the algorithm and in section 6 we conclude.

2 Preliminaries

In this work, an arm (any arbitrary one) is denoted as a_i and a^* is the optimal arm. The total time horizon is T. Any arbitrary round is denoted by m. The arm set containing all the arms is A, while $|A| = K$. Any arbitrary cluster is denoted by s_k. S_m denotes the set of all clusters $s_k \in S_m$ in the m-th round and $|S_m| = M \leq K$. The variable ℓ_m is the cluster size limit in each round m. $\hat{r}_i$ denotes the average estimated payoff of the i-th arm. The variable n_{s_k} denotes the number of times each arm $a_i \in s_k$ is pulled in each round m. The variable B_m denotes the arm set containing the arms that are not eliminated till round m. In any cluster $s_k \in S_m$, we denote $\hat{r}_{max_{s_k}} \in s_k$ as the maximum estimated payoff and $\hat{r}_{min_{s_k}} \in s_k$ as the minimum estimated payoff. $\Delta_i = r^* - r_i$ and also let $\Delta = min_{a_i \in A} \Delta_i$, that is it is the minimum possible gap over all arms in A.We also assume that all rewards are bounded in $[0, 1]$. For simplicity we will assume that there is only one optimal arm a^*.

3 Clustered UCB algorithm

Algorithm

The steps are presented in Algorithm 1.

Algorithm 1: ClusUCB

1: Pull each arm once and calculate $\hat{r}_i, \forall i \in A$
2: Let A be the set which contains all the arms
3: $B_0 := A$
4: $\ell_0 := 2, \epsilon_0 := 1, w_0 := 1$
5: For rounds $m = 0, 1, .. \lfloor \frac{1}{2} \log_2 \frac{T}{e} \rfloor$
6: Call SubroutineMerge
7: Pull each arm in s_k,
$n_m = \left\lceil \frac{2 \log (T\epsilon_m^2)}{\epsilon_m} \right\rceil$ number of times, $\forall s_k \in S_m$
8: **Arm Elimination:**
Delete all arms $a_i \in s_k$ for which
$\left\{ \hat{r}_i + \sqrt{\frac{\log (T\epsilon_m^2)}{2n_m}} \right\} < \max_{a_j \in s_k} \left\{ \hat{r}_j - \sqrt{\frac{\log (T\epsilon_m^2)}{2n_m}} \right\}$, $\forall s_k \in S_m$ and remove all such arms from B_m.
9: **Cluster Elimination:**
Delete all clusters $s_k \in S_m$ for which $\max_{a_i \in s_k} \left\{ \hat{r}_i + \sqrt{\frac{\rho \log (T\epsilon_m^2)}{2n_m}} \right\}$
$< \max_{a_j \in B_m} \left\{ \hat{r}_j - \sqrt{\frac{\rho \log (T\epsilon_m^2)}{2n_m}} \right\}$, where $\rho = \frac{1}{w_m}$ and remove all such arms in the cluster s_k from B_m to obtain B_{m+1}.
10: **Reset Parameters:**
11: $\epsilon_{m+1} := \frac{\epsilon_m}{2}$
12: $\ell_{m+1} := \min\{2\ell_m, K\}$
13: $w_{m+1} := 2w_m$
14: **Stopping Condition:**
Stop if $|B_m| = 1$ and pull $\max_{a_i \in B_m} \hat{r}_i$ till T is reached.

Subroutine: SubroutineMerge

Arrange all arms in B_m in ascending order based on their $\hat{r}_i$
Create clusters $s_i = \{a_i\}, \forall i \in B_m$
For $i = 1$ to K:
    For $j = i + 1$ to K:
        Merge s_i, s_j into s_i if $\exists a_i \in s_i$ and $\exists a_j \in s_j$, such that $|\hat{r}_i - \hat{r}_j| \leq \epsilon_m$ and $|s_i| + |s_j| \leq \ell_m$
Rename all clusters that have been formed as s_1 to s_M, where $|S_m| = M$ after merging

In the above algorithm we are dividing the arm set A into smaller clusters which belong to the cluster set S_m, where after merging $|S_m| = M \leq K$. We are bounding the cluster size by ℓ_m in each round, starting from an initial value of $\ell_0 = 2$ and doubling it after every round. For creating these clusters we rely on the parameter ϵ_m, which is initialized at 1 and halved after every round. Thus, we see that as the rounds progresses, ϵ_m gets smaller and smaller resulting in tighter and tighter clusters with high purity level. Single link clustering tends to create large clusters with many elements in a single cluster. To avoid this behavior, mainly because we are conducting exploration inside a cluster based on its size, we bound the maximum cluster size possible in any round by $\ell_m = \min\{2^m, K\}$. Hence as ϵ_m tends to Δ the upper limit on the cluster size gets fixed. We can also see that at any round m, the pulls n_m increases at a lesser rate compared to the round-based variants UCB-Revisited $\left\lceil \frac{2log(T\tilde{\Delta}_m^2)}{\tilde{\Delta}_m^2} \right\rceil$, where $\tilde{\Delta}_m$ is initialized at 1 and halved after every round or Median-Elimination $\frac{4}{\epsilon^2}\log(\frac{3}{\delta})$, where ϵ, δ are the parameters for PAC guarantee.

Next, in each of the round we eliminate arms like UCB-Revisited(Auer and Ortner (2010)) or Median Elimination(Even-Dar et al. (2006)),however, it is important to note that there is no longer a single point of reference based on which we are eliminating arms but now we have as many reference points to eliminate arms as number of clusters formed after merging that is $|S_m|$. We can be this aggressive because we have divided the larger problem into smaller sub-problems, doing local exploration and eliminating sub-optimal arms within each clusters with high guarantee. Hence, compared to UCB-Revisited or Median Elimination, the proposed algorithm should have a higher probability of arm deletion. Especially when K is large it is efficient to remove sub-optimal arms quickly rather getting tied down in hopeless exploration. Also our total regret depends on how many arms has survived till m-th round and so we don't need to keep track on the number of clusters formed.

Through cluster elimination condition we ensure that the stopping condition is reached faster. This is a much aggressive elimination condition and in the proofs we give a further analysis on why this works. The parameter $\rho \in (0, 1]$ in the confidence interval actually makes the cluster elimination a faster elimination condition than the elimination condition in UCB-Revisited(Auer and Ortner (2010)).

4 Main results

Here, we state the main theorem of the paper which shows the regret upper bound of ClusUCB.

Theorem 1. *Considering both the arm elimination and cluster elimination condition, the total regret till* T *is upper bounded by* $R_T \leq \sum_{i \in A:\Delta \geq b} \left\{ 3\Delta_i + \left(\frac{44}{\Delta_i^3}\right) + \left(\frac{2^{1+4\rho}\rho^{2\rho}T^{1-\rho}}{\Delta_i^{4\rho-1}}\right) + \left(\frac{32\log(T\frac{\Delta_i^4}{16})}{\Delta_i}\right) + \left(\frac{512\rho\log(T\frac{\Delta_i^4}{16\rho^2})}{\Delta_i}\right) \right\} + \sum_{i \in A:0 \leq \Delta_i \leq b} \left\{ \left(\frac{12}{b^3}\right) + \left(\frac{T^{1-\rho}2^{2\rho+\frac{3}{2}}}{\Delta_i^3}\right) + \left(\frac{T^{1-\rho}2^{2\rho+\frac{3}{2}}}{b^{4\rho-1}}\right) \right\} + max_{i:\Delta \leq b}\Delta_i T,$

where $\rho \in (0, 1]$ and T is the horizon.

Proof. See Appendix. □

Remark 1. *Thus, we see the most significant term in the regret is $\frac{512\rho \log (T\frac{\Delta_i^4}{12})}{\Delta_i}$, which is significantly lower than UCB1, UCB2, EXP3, UCB(δ), MOSS, UCB-Revisited and Median Elimination under certain cases when the $\Delta \to 0$ and K is large. Also ClusUCB is more efficient than UCB1, MOSS, EXP3, UCB(δ), KL-UCB as K scales up because being a round-based algorithm, it is removing sub-optimal arms in each round as opposed to the former algorithms which are calculating the confidence bounds over all arms in each timestep and then choosing the max of them.*

Remark 2. *A sketch of the proof is given here. In the first step, we try to calculate the number of pulls n_m required to make the optimal arm safe with a high probability so that it goes up atleast $\frac{\hat{\Delta}_{s_i}}{2}$ where $\hat{\Delta}_{s_i} = max_{i \in s_i} \hat{r}_i - min_{j \in s_i} \hat{r}_j, i \neq j$. This is shown in Proposition 1. Second step, we try to bound the probability of arm elimination of any sub-optimal arm without cluster elimination. This is shown in Proposition 2. Third step, we try to bound the probability of cluster elimination(without any arm elimination) with all arms within it and the favourable event leading to it. This is shown in Proposition 3. Finally, in the proof of Theorem 1 we combine all these to get the regret bound.*

Proposition 1. *The probability that the optimal arm $a^* \in s_i$ will lie above $\hat{r}_{min_{s_i}} + \frac{\hat{\Delta}_{s_i}}{2}$ after $\left\lceil \frac{2\log(\psi(m)T\epsilon_m^2)}{\epsilon_m} \right\rceil$ pulls in the m-th round is given by $\left\{1 - \frac{1}{(\psi(m)T\epsilon_m^2)^{\ell_m^2 \epsilon_m}}\right\}$ where $\hat{r}_{min_{s_i}}$ is the minimum payoff in s_i, $\hat{\Delta}_{s_i} = max_{i \in s_i} \hat{r}_i - min_{j \in s_i} \hat{r}_j, i \neq j$, ϵ_m is halved after every round and T is the horizon.*

The proof of Proposition 1 is given in **Appendix A**.(Supplementary material)

Proposition 2. *Considering only the arm elimination condition, the total regret till T is upper bounded by $R_T \leq \sum_{i \in A: \Delta_i \geq b} \left\{ \left(\frac{44}{(\Delta_i)^3} \right) + \left(\Delta_i + \frac{32 \log (T\frac{\Delta_i^4}{16})}{\Delta_i} \right) \right\} + \sum_{i \in A: 0 \leq \Delta_i \leq b} \frac{12}{b^3} + max_{i: \Delta \leq b} \Delta_i T$, where T is the horizon.*

The proof of Proposition 2 is given in **Appendix B**.(Supplementary material).

Remark 3. *Thus, we see that the confidence interval term $c_m = \sqrt{\frac{\log(T\epsilon_m^2)}{2n_m}}$ makes the algorithm eliminate an arm a_i as soon as $\sqrt{\epsilon_m} < \frac{\Delta_i}{2}$. The above result is in contrast with UCB-Revisited which only deletes an arm if $\tilde{\Delta}_m < \frac{\Delta_i}{2}$, where $\tilde{\Delta}_m$ is initialized at 1 and is halved after every round. We also see from our result that a much less stricter elimination condition is adopted inside a cluster. We are eliminating inside a cluster because we are exploring locally and we are guaranteed with $\left\{1 - \frac{2}{(T\epsilon_m^2)^{\ell_m^2 \epsilon_m}}\right\}$ probability that in the m-th round optimal arm a^* will atleast lie above $\hat{r}_{min_{s_i}} + \hat{\Delta}_{s_i}$.*

Proposition 3. *Considering only the cluster elimination condition, the total regret till T is upper bounded by $R_T \leq \sum_{i \in A: \Delta_i \geq b} \left(\frac{2^{1+4\rho} \rho^{2\rho} T^{1-\rho}}{\Delta_i^{4\rho-1}} \right) + \left(\Delta_i + \frac{512\rho \log (T\frac{\Delta_i^4}{16\rho^2})}{\Delta_i} \right) + \left(\frac{T^{1-\rho} 2^{2\rho+\frac{3}{2}}}{\Delta_i^{4\rho-1}} \right) \Big\} + \sum_{i \in A: 0 \leq \Delta_i \leq b} \left(\frac{T^{1-\rho} 2^{2\rho+\frac{3}{2}}}{b^{4\rho-1}} \right) + max_{i: \Delta \leq b} \Delta_i T$, where $\rho \in (0, 1)$ and T is the horizon.*

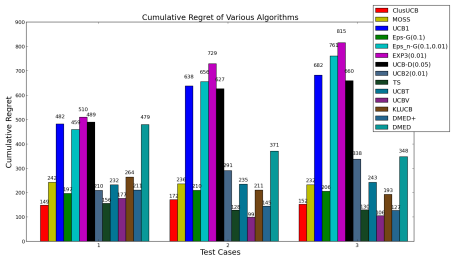

Figure 1: Experiment 1: Regret for various Algorithms in 3 testcases. $T = 20000; \psi(m) = 1.5/m$

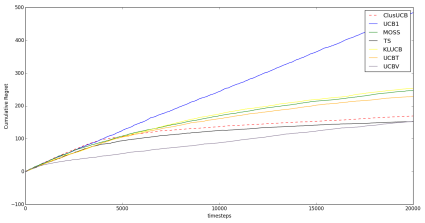

Figure 2: Experiment 2: Growth of Regret for test case 1. $T = 20000; \psi(m) = 1.5/m$

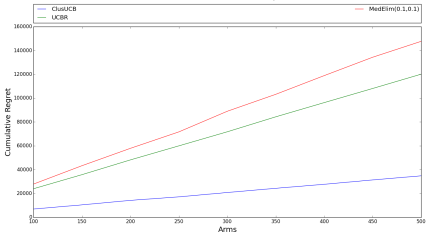

Figure 3: Experiment 3: Regret for ClusUCB, UCB-Revisited and Median-Elimination. $T = 5 \times 10^5; \psi(m) = 0.1/m$

The proof of Proposition 3 is given in **Appendix C**.(Supplementary material)

5 Simulation experiments

The first experiment is conducted over a testbed of 10 arms for the 3 test-cases involving Bernoulli reward distribution with expected rewards of the arms as described below.

Test Case 1: [0.07, 0.07, 0.07, 0.07, 0.07, 0.07, 0.07, 0.07, 0.07, 0.1]
Test Case 2: [0.07, 0.07, 0.07, 0.05, 0.05, 0.05, 0.05, 0.05, 0.05, 0.1]
Test Case 3: [0.07, 0.07, 0.07, 0.05, 0.05, 0.05, 0.03, 0.03, 0.03, 0.1]

These type of cases are frequently encountered in web-advertising domain. The regret is averaged over 50 independent runs over each testbed and is shown in **Fig 1**. 14 algorithms ClusUCB, MOSS, UCB1, UCB2($\alpha = 0.01$), ϵ-greedy($\epsilon = 0.1$), ϵ_n-greedy($c = 0.1, d = 0.01$), Exp3($\gamma = 0.01$), UCB-Delta($\delta = 0.05$), UCB-Tuned, UCB-V, KL-UCB, DMED+, DMED(as stated in Garivier and Cappé (2011)) and Thompson Sampling are run over this testbed and shown in this figure(the cumulative regret averaged over 50 independent runs is shown above each bar). Here, we see that except Thompson Sampling and UCB-V and DMED+ the regret of Clustered-UCB is lower than the rest for all the test cases. Even in test case 1, the regret of ClusUCB is nearly same as Thompson Sampling and better than UCB-V and DMED+. In test case 1 the regret is so low for ClusUCB because $\Delta_i = \Delta, \forall i \in A$ and while all the algorithms employ significant exploration ClusUCB by virtue of dividing the problem into sub-problems quickly finds the optimal arm. The parameters of ϵ_n-greedy are very difficult to estimate and if d is not a tight lower bound of Δ then the result can be poor as shown in the figure. The parameter less algorithms UCB1, MOSS, UCB-Tuned, Thompson Sampling and UCB-V are run as mentioned in the respective papers. For algorithms requiring parameters, such as UCB2, UCB-Delta, ϵ-greedy, ϵ_n-greedy, EXP3, the parameters were tried and tested over several values and then implemented in each of the test cases. KL-UCB, DMED and DMED+ code

is taken from Cappe et al. (2012), and is run accordingly the way author specified with KL-UCB parameter $c = 0$. KL-UCB regret is also poorer than ClusUCB in test case 1 and 2 and same as ClusUCB in test case 3. It takes significant more time to run the algorithm than ClusUCB which might not be feasible in many real world scenarios like web advertising. In this short horizon for all the test cases ClusUCB performs better than MOSS and UCB1. DMED+, UCB-V and TS, as expected beats ClusUCB in scenario 2 and 3. In test cases 2 and 3, ClusUCB performs bad mainly because here the Δ_i is much more evenly spread and ClusUCB takes significantly more exploration to find the optimal arm. Also since it's a short horizon of $T = 20000$, we have taken $\psi(m) = \dfrac{1.5}{m}$.

The second experiment is conducted over the same testbed as in test case 1 above and shown in **Fig 2**. We check the growth of regret over time for 7 algorithms as mentioned in the figure. Here, we see that ClusUCB has a much steeper regret curve than the rest which signifies a faster exploitation and less exploration. It quickly finds the optimal arm and the cumulative regret nearly becomes negligible. UCB1 is not able to find the optimal arm in this short horizon whereas UCB-V performs much better but its regret still does not stabilize within this short horizon. TS performs well as well whereas MOSS performs much worse and hardly stabilizes in this short horizon. We also see that ClusUCB is remarkably stable in this short horizon and it outputed a sub-optimal arm only 5 out of 50 runs. Comparing this with UCB-V, it outputed a wrong arm 11 out of 50 runs.

The third experiment is conducted over a testbed of $100 - 500$ arms at an interval of 50 arms, of which (for any particular run) $\frac{1}{3}$ arms have Gaussian reward distribution $N(\mu = 0.2, \sigma = 0.3)$, rest $\frac{2}{3}$ arms have Gaussian reward distribution $N(\mu = 0.7, \sigma = 0.3)$ and the optimal arm have parameters $N(\mu = 0.9, \sigma = 0.3)$. We conduct this experiment to not only check the performance of ClusUCB in Gaussian distribution, but also the growth of regret over a large set of arms. Here, we employ only select algorithms(round based algorithms) since the action space is very large and the other algorithms will take a long time to converge on the optimal arm. The regret is averaged over 50 independent runs over this testbed and is shown in **Fig 3**. Three round-based algorithms ClusUCB, UCB-Revisited and Median Elimination are run over this testbed and shown in the figure. Here, Median-Elimination performs worse than UCB-Revisited. We also see that over this large set of arms the regret of ClusUCB is not only much lesser than UCB-Revisited and Median-Elimination but also being a round based algorithm it is much faster than other standard algorithms. Again since it's a large horizon of $T = 5 \times 10^5$ so we have taken $\psi(m) = \dfrac{0.1}{m}$.

6 Conclusions and future work

Our study concludes that the regret of ClusUCB is lower than UCB1, UCB2, EXP3, MOSS, UCB-Revisited, KL-UCB, UCB-Tuned, DMED and Median Elimination under certain cases when the $\Delta \to 0$ and K is large. Such cases can frequently occur in web advertising scenarios where the r_i are small and a large number of ads are available in the pool. We see that the upper bound of ClusUCB has $\log$ dependence on horizon T whereas MOSS has a dependence of $T^{1/2}$ in the distribution free case. For the critical case when $\Delta_{i:r_i<r*}, \forall i \in A$ are equal then ClusUCB performs better than UCB-Revisited, UCB1, UCB-Tuned, KL-UCB, MOSS and EXP3 with sufficient tuning of parameters as shown empirically. Also ClusUCB scales well with large K as compared with UCB1, EXP3, MOSS, KL-UCB and Thompson Sampling since it is eliminating sub-optimal arms after some timesteps. We must also remember as the number of arms increases UCB1, UCB2, KL-UCB and UCB(δ) will take more time as all of these algorithms have to build their confidence set over all the arms which will take $O(K)$ time whereas algorithms like ClusUCB, UCB-Revisited and Median Elimination will take much lesser time as they keep on removing sub-optimal arm after each round. KL-UCB because of its calculation of the divergence function performs much slower as compared to ClusUCB. Also, ClusUCB can be used in the budgeted bandit setup since within a fixed

horizon/budget T, the algorithm comes up with an optimal arm with an exponentially decreasing error probability as proved theoretically in this work. Further uses of this algorithm can be in the contextual bandit scenario whereby the clustering of arms can be done on the basis of the feature vectors of the arms and users.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320, 2011.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. *arXiv preprint arXiv:1111.1797*, 2011.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, pages 217–226, 2009.

Jean-Yves Audibert and Sébastien Bubeck. Best arm identification in multi-armed bandits. In *COLT-23th Conference on Learning Theory-2010*, pages 13–p, 2010.

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

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

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

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.

Olivier Cappe, Aurelien Garivier, and Emilie Kaufmann. pymabandits, 2012. `http://mloss.org/software/view/415/`.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *The Journal of Machine Learning Research*, 7: 1079–1105, 2006.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. *arXiv preprint arXiv:1102.2490*, 2011.

Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *COLT*, pages 67–79. Citeseer, 2010.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985.

Herbert Robbins. Some aspects of the sequential design of experiments. In *Herbert Robbins Selected Papers*, pages 169–177. Springer, 1952.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, pages 285–294, 1933.

Appendix

A Appendix A

Proposition 4. *The probability that the optimal arm $a^* \in s_i$ will lie above $\hat{r}_{min_{s_i}} + \frac{\hat{\Delta}_{s_i}}{2}$ after $\left\lceil \frac{2\log(T\epsilon_m^2)}{\epsilon_m} \right\rceil$ pulls in the m-th round is given by $\left\{1 - \frac{1}{(T\epsilon_m^2)^{\ell_m^2 \epsilon_m}}\right\}$ where $\hat{r}_{min_{s_i}}$ is the minimum payoff in s_i, $\hat{\Delta}_{s_i} = max_{i \in s_i}\hat{r}_i - min_{j \in s_i}\hat{r}_j, i \neq j$, ϵ_m is halved after every round and T is the horizon.*

Proof. of Proposition 1:
We start by considering the worst case scenario that in the m-th round, in a cluster s_i, the optimal arm a^* has performed worst, such that $\hat{r}^* < \hat{r}_i, \forall a_i \in s_i$. Let, $\hat{\Delta}_{s_i} = max_{i \in s_i}\hat{r}_i - min_{j \in s_i}\hat{r}_j$ where $i \neq j$. Also, let $|s_i| = k_{s_i}$ and $\hat{r}^* = \hat{r}_{min_{s_i}} \leq \hat{r}_i, \forall i \in s_i$ also $\hat{r}_{max_{s_i}} \geq \hat{r}_i, \forall i \in s_i$.
Again, given that there are k_{s_i} number of arms in s_i, and for each $a_i, a_j \in s_i$ since $|\hat{r}_i - \hat{r}_j| \leq \epsilon_m$, the longest possible gap is $(k_{s_i} - 1)\epsilon_m$ which is greater than the actual estimated gap $\hat{\Delta}_{s_i}$.
So, $\hat{\Delta}_{s_i} \leq (k_{s_i} - 1)\epsilon_m$, as $|\hat{r}_i - \hat{r}_j| \leq \epsilon_m, \forall i, j \in s_i$
$\leq \ell_m \epsilon_m$, as $k_{s_i} \leq \ell_m$
Now, applying Chernoff-Hoeffding bound and considering independence of events,
$\mathbb{P}\{\hat{r}^* \leq r^* + \frac{\hat{\Delta}_{s_i}}{2}\} \Rightarrow \mathbb{P}\{\hat{r}^* \leq r^* + \frac{\ell_m \epsilon_m}{2}\} \leq exp(-2\frac{(\ell_m \epsilon_m)^2}{4} n^*)$
Now, putting $n_m = n^* = \frac{2\log(T\epsilon_m^2)}{\epsilon_m}$
$\mathbb{P}\{\hat{r}^* \leq r^* + \frac{\ell_m \epsilon_m}{2}\} \leq exp(-\ell_m^2 \epsilon_m \log(T\epsilon_m^2))$
$\mathbb{P}\{\hat{r}^* \leq r^* + \frac{\ell_m \epsilon_m}{2}\} \leq \frac{1}{(T\epsilon_m^2)^{\ell_m^2 \epsilon_m}}$
Hence, the probability that the optimal arm a^* after n_m pulls going above $\hat{r}_{min_{s_i}} + \frac{\hat{\Delta}_{s_i}}{2}$ is $\left\{1 - \frac{1}{(T\epsilon_m^2)^{\ell_m^2 \epsilon_m}}\right\}$ □

B Appendix B

Proposition 5. *Considering only the arm elimination condition, the total regret till T is upper bounded by $R_T \leq \sum_{i \in A: \Delta_i \geq b} \left\{ \left(\frac{44}{(\Delta_i)^3} \right) + \left(\Delta_i + \frac{32 \log (T \frac{\Delta_i^4}{16})}{\Delta_i} \right) \right\} + \sum_{i \in A: 0 \leq \Delta_i \leq b} \frac{12}{b^3} + max_{i: \Delta \leq b} \Delta_i T$, where T is the horizon.*

Proof. of Proposition 5:

Let, for each sub-optimal arm a_i, $m_i = \min\{m | \sqrt{\epsilon_m} \leq \frac{\Delta_i}{2}\}$ be the first round when $\sqrt{\epsilon_m} \leq \frac{\Delta_i}{2}$.

B.1 Case a:

Some sub-optimal arm a_i is not eliminated in round m_i or before and the optimal arm $a^* \in B_{m_i}$
In arm elimination condition, given the choice of confidence interval c_m, we want to bound the event $\hat{r}_i + c_{m_i} \leq \hat{r}^* - c_{m_i}$.

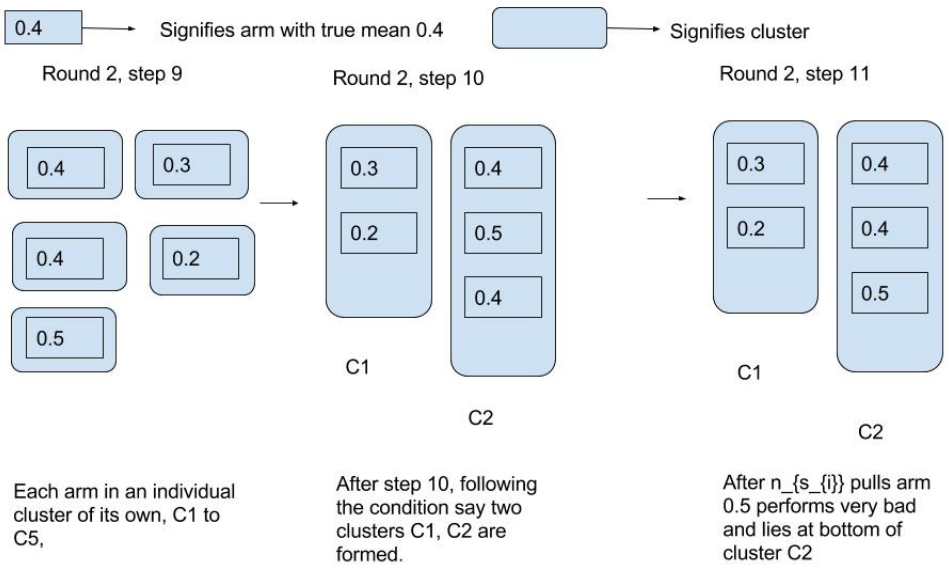

Figure 4: Steps 9-11

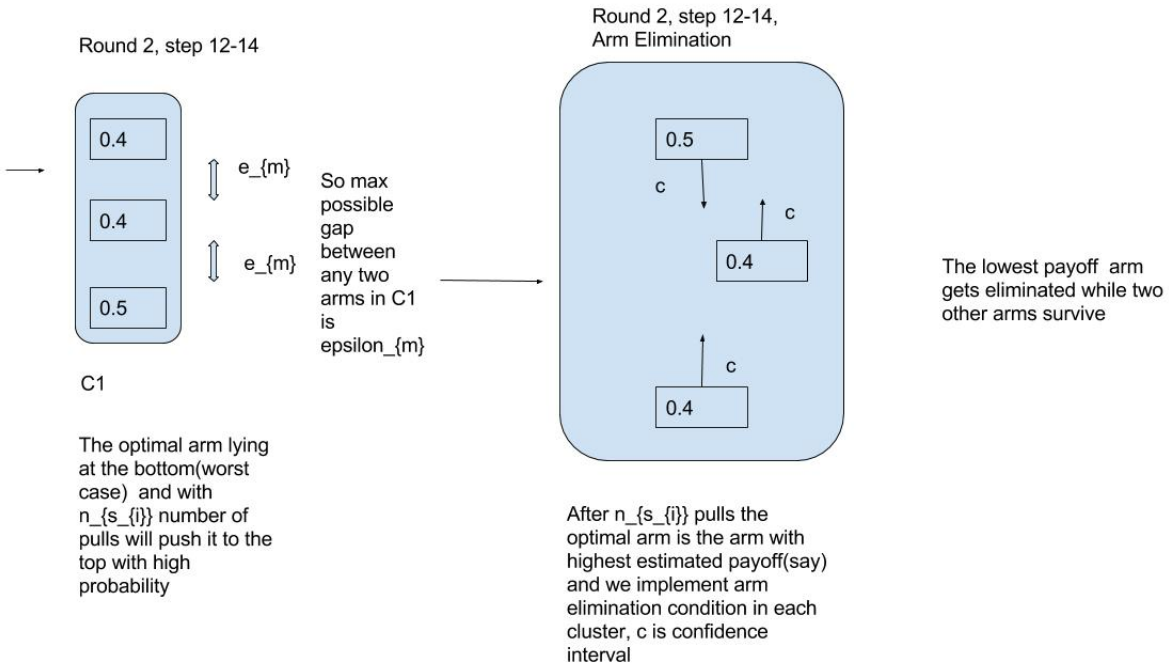

Figure 5: Steps 12-14

Now, $c_{m_i} = \sqrt{\frac{\log(T\epsilon_{m_i}^2)}{2n_{m_i}}}$.

Putting the value of $n_{m_i} = \frac{2\log(T\epsilon_{m_i}^2)}{\epsilon_{m_i}}$ in c_{m_i},

$c_{m_i} = \sqrt{\frac{\epsilon_{m_i}\log(T\epsilon_{m_i}^2)}{2*2\log(T\epsilon_{m_i}^2)}} = \frac{\sqrt{\epsilon_{m_i}}}{2} = \sqrt{\epsilon_{m_i+1}} < \frac{\Delta_i}{4}$

Again, $\exists a_i \in s_i$ such that, $\hat{r}_i + c_{m_i} \leq r_i + 2c_{m_i}$

$$\begin{aligned} &= \hat{r}_i + 4c_{m_i} - 2c_{m_i} \\ &\leq r_i + \Delta_i - 2c_{m_i} \\ &< r^* - 2c_{m_i} \\ &\leq \hat{r}^* - c_{m_i} \end{aligned}$$

Hence, we get that as soon as $\sqrt{\epsilon_{m_i}} < \frac{\Delta_i}{2}$, $\exists a_i$ which gets eliminated.

So, we need to bound the event of $\hat{r}_i + c_{m_i} \leq \hat{r}^* - c_{m_i}$ given that $\sqrt{\epsilon_{m_i}} < \frac{\Delta_i}{2}$ becomes true for some arm a_i after the m-th round and $c_m = \sqrt{\frac{\log(T\epsilon_{m_i}^2)}{2n_{m_i}}}$.

So, we need to bound the probability,

$\mathbb{P}\{\hat{r}^* \leq r^* - c_{m_i}\} \leq U_m$, where U_m is an arbitrary upper bound.

Applying Chernoff-Hoeffding bound and considering independence of events,

$$\begin{aligned} \mathbb{P}\{\hat{r}^* \leq r^* - c_{m_i}\} &\leq exp(-2c_{m_i}^2 n_{m_i}) \\ &\leq exp(-2 * \frac{\log(T\epsilon_{m_i}^2)}{2n_{m_i}} * n_{m_i}) \\ &\leq \frac{1}{T\epsilon_{m_i}^2} \end{aligned}$$

Similarly, $\mathbb{P}\{\hat{r}_i \geq r_i + c_{m_i}\} \leq \frac{1}{T\epsilon_{m_i}^2}$

Summing, the two up, the probability that a sub-optimal arm a_i is not eliminated in m_i-th round is $\left(\frac{2}{T\epsilon_{m_i}^2}\right)$.

Summing up over all arms in A and bounding trivially by $T\Delta_i$,

$$\sum_{i\in A}\left(\frac{2T\Delta_i}{T\epsilon_{m_i}\frac{\Delta_i}{2}^4}\right) \leq \sum_{i\in A}\left(\frac{8}{\epsilon_{m_i}\Delta_i}\right) \leq \sum_{i\in A}\left(\frac{32}{\Delta_i^3}\right)$$

B.2 Case b1:

Either an arm a_i is eliminated in round m_i or before or else there is no optimal arm $a^* \in B_{m_i}$.

Also, since we are eliminating a sub-optimal arm a_i on or before round m_i, it is pulled no longer than,

$$n_{m_i} = \left\lceil \frac{2\log(T\epsilon_{m_i}^2)}{\epsilon_{m_i}} \right\rceil$$

So, the total contribution of a_i till round m_i is given by,

$$\Delta_i \left\lceil \frac{2\log(T\epsilon_{m_i}^2)}{\epsilon_{m_i}} \right\rceil \leq \Delta_i \left\lceil \frac{2\log(T(\frac{\Delta_i}{2})^4)}{(\frac{\Delta_i}{2})^2} \right\rceil, \text{ since } \sqrt{\epsilon_{m_i}} \leq \frac{\Delta_i}{2}$$

$$\leq \Delta_i \left(1 + \frac{32 \log (T(\frac{\Delta_i}{2})^4)}{\Delta_i^2}\right)$$
$$\leq \Delta_i \left(1 + \frac{32 \log (T\frac{\Delta_i^4}{16})}{\Delta_i^2}\right)$$

Summing over all arms,

$$\leq \sum_{i \in A} \Delta_i \left(1 + \frac{32 \log (T\frac{\Delta_i^4}{16})}{\Delta_i^2}\right)$$

B.3 case b2:

In this case we will consider that the optimal arm a^* was eliminated by a sub-optimal arm. Firstly, if conditions of case b1 holds then the optimal arm a^* will not be eliminated in round $m = m_*$ or it will lead to the contradiction that $r_i > r^*$. In any round m_*, if the optimal arm a^* gets eliminated then for any round from 1 to m_j all arms a_j such that $\sqrt{\epsilon_m} < \frac{\Delta_j}{2}$ were eliminated according to assumption in case b1. Let, the arms surviving till m_* round be denoted by $A^{'}$. This leaves any arm a_b such that $\sqrt{\epsilon_m} \geq \frac{\Delta_b}{2}$ to still survive and eliminate arm a^* in round m_*. Let, such arms that survive a^* belong to $A^{''}$. Also maximal regret per step after eliminating a^* is the maximal Δ_j among the remaining arms a_j with $m_j \geq m_*$. Let m_b be the round when $\sqrt{\epsilon_m} < \frac{\Delta_b}{2}$ that is $m_b = min\{m | \sqrt{\epsilon_m} < \frac{\Delta_b}{2}\}$. Hence, the maximal regret after eliminating the arm a^* is upper bounded by,

$\sum_{m_*=0}^{max_{j \in A'} m_j} \sum_{i \in A'': m_i > m_*} \left(\frac{2}{T\epsilon_m^2}\right).T \max_{j \in A'': m_j \geq m_*} \Delta_j$

$\leq \sum_{m_*=0}^{max_{j \in A'} m_j} \sum_{i \in A'': m_i > m_*} \left(\frac{2}{T\epsilon_m^2}\right).T.2\sqrt{\epsilon_m}$, since $\sqrt{\epsilon_m} < \frac{\Delta_i}{2}$

$\leq \sum_{m_*=0}^{max_{j \in A'} m_j} \sum_{i \in A'': m_i > m_*} 4\left(\frac{1}{\epsilon_m^{3/2}}\right)$

$\leq \sum_{i \in A'': m_i > m_*} \sum_{m_*=0}^{\min\{m_i, m_b\}} \left(\frac{4}{2^{-(3/2)m_*}}\right)$

$\leq \sum_{i \in A'} \left(\frac{4}{2^{-(3/2)m_*}}\right) + \sum_{i \in A'' \setminus A'} \left(\frac{4}{2^{-(3/2)m_b}}\right)$

$\leq \sum_{i \in A'} \left(\frac{4 * 2^{3/2}}{\Delta_i^3}\right) + \sum_{i \in A'' \setminus A'} \left(\frac{4 * 2^{3/2}}{b^3}\right)$

$\leq \sum_{i \in A'} \left(\frac{12}{\Delta_i^3}\right) + \sum_{i \in A'' \setminus A'} \left(\frac{12}{b^3}\right)$

Summing up **Case a**, **Case b1** and **Case b2**, the total regret till round m is given by,

$$R_T \leq \sum_{i \in A: \Delta_i \geq b} \left\{ \left(\frac{44}{(\Delta_i)^3}\right) + \left(\Delta_i + \frac{32 \log (T\frac{\Delta_i^4}{16})}{\Delta_i}\right) \right\} + \sum_{i \in A: 0 \leq \Delta_i \leq b} \frac{12}{b^3} + max_{i: \Delta \leq b} \Delta_i T \qquad \square$$

C Appendix C

An illustrative diagram explaining Cluster Elimination is given in Figure 6.

Proposition 6. *Considering only the cluster elimination condition, the total regret till T is upper bounded by*

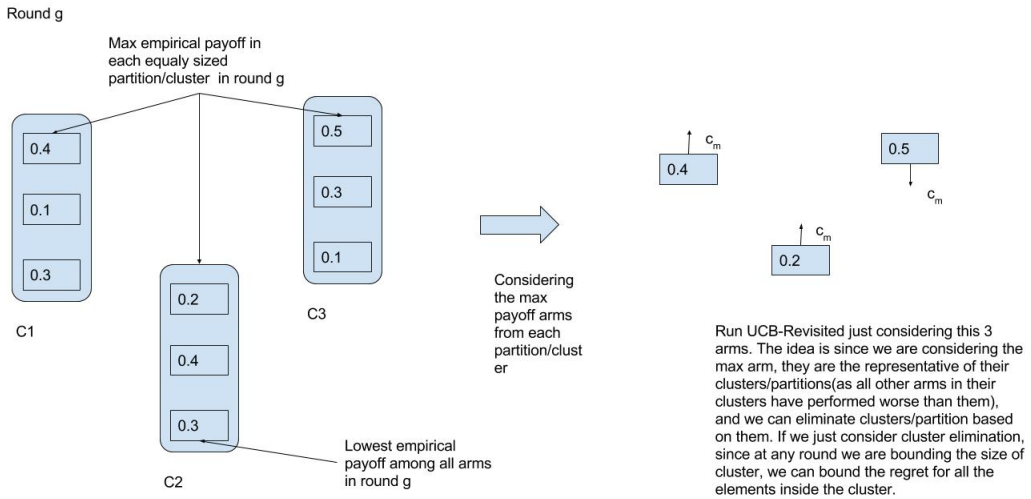

Figure 6: Cluster Elimination

$$R_T \leq \sum_{i\in A:\Delta_i\geq b}\left(\frac{2^{1+4\rho}\rho^{2\rho}T^{1-\rho}}{\Delta_i^{4\rho-1}}\right)+\left(\Delta_i+\frac{512\rho\log\left(T\frac{\Delta_i^4}{16\rho^2}\right)}{\Delta_i}\right)+\left(\frac{T^{1-\rho}2^{2\rho+\frac{3}{2}}}{\Delta_i^{4\rho-1}}\right)\Bigg\}+\sum_{i\in A:0\leq\Delta_i\leq b}\left(\frac{T^{1-\rho}2^{2\rho+\frac{3}{2}}}{b^{4\rho-1}}\right)+$$
$max_{i:\Delta\leq b}\Delta_i T$, *where* $\rho\in(0,1)$ *and* T *is the horizon.*

Remark 4. *A sketch of the proof is given below,*

- *Define* g*-th round(in proposition 6) as the same way as the* m*-th round signifying the first/minimum round when a cluster gets eliminated.*
- *Let in the* g*-th round* C_g *be the set which contains all the max payoff arms from each cluster.*
- *For regret bound proof according to proposition* 5*, consider only this* C_g *and proof following the same way as in proposition* 5 *with some minor change. Hence,* C_g *actually behaves like the set* B_m *in proposition* 5 *but contains just the max payoff arm from each cluster.*
- *Introduce a bounded parameter* $\rho\in(0,1]$ *for cluster elimination to mimic the arm elimination condition in proposition* 5*, but with the condition that whenever* $\sqrt{\rho\epsilon_g}\leq\frac{\Delta_i}{2}, a_i\in C_g$*, then the cluster* s_k*(where* $\hat{r}_i$ *is the max payoff) gets eliminated. Because of the algorithm, we are guaranteed that the size of the cluster in the* g*-th round is* $\ell_g=\min\{2^g,K\}$ *and the maximum gap within a cluster that is* $\hat{\Delta}_{s_k}=(\hat{r}_{max_{s_k}}-\hat{r}_{min_{s_k}})\leq\ell_g\epsilon_g$*, where* $\epsilon_g=2^{-g}$*.*

Remark 5. *We also point out a few other observations:*

- *The maximum number of clusters that can be formed in any round is* $\lceil\frac{K}{2}\rceil$*, since we start with a minimum cluster size of* $\ell_g=2$*.*
- *Also as* ϵ_g *decreases after every round and as* $\epsilon_g\to\Delta, |S_g|\to D$*, where* D *is the true number of clusters based on the underlying distribution of* $r_i, \forall i\in A$*.*

- *On the implementation side, the subroutine SubroutineMerge is implemented in such a way, that after arranging the arms in ascending order based on their $\hat{r}_i$, a sweep from left to right will put the arms in their respective clusters in such a way that $|\hat{r}_i - \hat{r}_j|, \forall i, j \in B_g$ is at most ϵ_g.*

Proof. of Proposition 6:

Let $C_g = \{\hat{r}_{max_{s_i}} | \forall s_i \in S\}$, that is let C_g be the set of all arms which has the maximum estimated payoff in their respective clusters in the g-th round.

Let, for each sub-optimal arm $a_i \in C_g$, $g_i = \min\{g | \sqrt{\rho\epsilon_g} \leq \frac{\Delta_i}{2}\}$. So, g_i be the first round when $\sqrt{\rho\epsilon_g} \leq \frac{\Delta_i}{2}$ where $a_i \in C_g$ is the maximum payoff arm in cluster s_k. We will also consider that $\max \hat{r}_i \in C_g$ is a^* and it has not still been eliminated. The parameter ρ is introduced just to make sure that the cluster elimination is a more aggressive elimination than arm elimination.

So, for cluster elimination we will only be considering the arms in C_g and following the proof of proposition 5,

C.1 Case a:

Some sub-optimal arm a_i is not eliminated in round g_i or before and the optimal arm $a^* \in B_{g_i} \subset B_{m_i}$
In arm elimination condition, given the choice of confidence interval c_g, we want to bound the event $\hat{r}_i + c_{g_i} \leq \hat{r}^* - c_{g_i}$.

Now, $c_{g_i} = \sqrt{\frac{\rho \log(T\epsilon_{g_i}^2)}{2n_{g_i}}}$, where $0 < \rho \leq 1$

Putting the value of $n_{g_i} = \frac{2\log(T\epsilon_{g_i}^2)}{\epsilon_{g_i}}$ in c_{g_i},

$$c_{g_i} = \sqrt{\frac{\rho * \epsilon_{g_i} \log(T\epsilon_{g_i}^2)}{2 * 2\log(T\epsilon_{g_i}^2)}} = \sqrt{\frac{\rho\epsilon_{g_i}}{2}} = \sqrt{\rho\epsilon_{g_i+1}} < \frac{\sqrt{\rho}\Delta_i}{4} < \frac{\Delta_i}{4}$$

Again, $\exists a_i \in C_g$ such that, $\hat{r}_i + c_{g_i} \leq r_i + 2c_{g_i}$

$$\begin{aligned} &= \hat{r}_i + 4c_{g_i} - 2c_{g_i} \\ &\leq r_i + \Delta_i - 2c_{g_i} \\ &< r^* - 2c_{g_i} \\ &\leq \hat{r}^* - c_{g_i} \end{aligned}$$

Hence, we get that as soon as $\sqrt{\rho\epsilon_{g_i}} < \frac{\Delta_i}{2}$, $\exists a_i \in C_g$ which gets eliminated.

So, we need to bound the event of $\hat{r}_i + c_{g_i} \leq \hat{r}^* - c_{g_i}$ given that $\sqrt{\rho\epsilon_{g_i}} < \frac{\Delta_i}{2}$ becomes true for some arm $a_i \in C_g$ after the g-th round and $c_{g_i} = \sqrt{\frac{\rho \log(T\epsilon_{g_i}^2)}{2n_{g_i}}}$.

So, we need to bound the probability,

$\mathbb{P}\{\hat{r}^* \leq r^* - c_{g_i}\} \leq U_g$, where U_g is an arbitrary upper bound.

Applying Chernoff-Hoeffding bound and considering independence of events,

$$\begin{aligned} \mathbb{P}\{\hat{r}^* \leq r^* - c_{g_i}\} &\leq exp(-2c_{g_i}^2 n_{g_i}) \\ &\leq exp(-2 * \frac{\rho \log(T\epsilon_{g_i}^2)}{2n_{g_i}} * n_{g_i}) \\ &\leq \frac{1}{(T\epsilon_{g_i}^2)^\rho} \end{aligned}$$

Similarly, $\mathbb{P}\{\hat{r}_i \geq r_i + c_{g_i}\} \leq \frac{1}{(T\epsilon_{g_i}^2)^\rho}$

Summing, the two up, the probability that a sub-optimal arm $a_i \in C_g$ is not eliminated in g_i-th round is $\left(\frac{2}{(T\epsilon_{g_i}^2)^\rho}\right)$.

Now, for each round g, all the elements of C_g are the respective max payoff arms of their cluster $s_k, \forall s_k \in S_g$, that is all the other arms in their respective clusters have performed worse than them. Hence, since $A \supset C_g$ and we can bound the max probability that a sub-optimal arm $a_j \in A$ is not eliminated in the g-th round by the same probability of $\left(\frac{2}{(T\epsilon_{g_i}^2)^\rho}\right)$.

Summing up over all arms in A and bounding trivially by $T\Delta_i$,

$$\sum_{i\in A}\left(\frac{2T\Delta_i}{(T\frac{\Delta_i}{16\rho^2}^4)^\rho}\right) \leq \sum_{i\in A}\left(\frac{2^{1+4\rho}T^{1-\rho}\rho^{2\rho}\Delta_i}{\Delta_i^{4\rho}}\right)$$

$$\leq \sum_{i\in A}\left(\frac{2^{1+4\rho}\rho^{2\rho}T^{1-\rho}}{\Delta_i^{4\rho-1}}\right)$$

Thus, we see that putting a value of $\rho = 1$, nicely brings out the result of case b1 of proposition 5.

C.2 Case b1:

Either an arm $a_i \in C_g$ is eliminated along with all the arms in that cluster in round g_i(or before) or else there is no optimal arm $a^* \in C_{g_i}$. Again, in the g-th round, the maximum total elements in the cluster can be no more than $\ell_{g_i} = \min\{2^g, K\}$.

Also, since we are eliminating a sub-optimal arm $a_i \in C_{g_i}$ on or before round g_i, it is pulled (along with all the other arms in that cluster) no longer than,

$$n_{g_i} = \left\lceil \frac{2\log(T\epsilon_{g_i}^2)}{\epsilon_{g_i}} \right\rceil$$

So, the total contribution of a_i along with all the other arms in the cluster till round g_i is given by,

$$\Delta_i\left\lceil \frac{2\ell_{g_i}\log(T\epsilon_{g_i}^2)}{\epsilon_{g_i}} \right\rceil \leq \Delta_i\left\lceil \frac{2\ell_{g_i}\log(T(\frac{\Delta_i}{2\sqrt{\rho}})^4)}{(\frac{\Delta_i}{2\sqrt{\rho}})^2} \right\rceil, \text{ since } \sqrt{\rho\epsilon_{g_i}} \leq \frac{\Delta_i}{2}$$

$$\leq \Delta_i\left(1 + \frac{32 * 2^{g_i} * \rho * \log(T(\frac{\Delta_i}{2\sqrt{\rho}})^4)}{\Delta_i^2}\right), \text{ since in the } g\text{-th round the maximum cluster size is bounded by } \min\{2^g, K\}.$$

$$\leq \Delta_i\left(1 + \frac{32 * 16 * \rho * \log(T(\frac{\Delta_i}{2\sqrt{\rho}})^4)}{\Delta_i^2}\right)$$

$$\leq \Delta_i\left(1 + \frac{512\rho\log(T\frac{\Delta_i^4}{16\rho^2})}{\Delta_i^2}\right)$$

Summing over all arms in $A \supset C_g$,

$$\leq \sum_{i\in A}\Delta_i\left(1 + \frac{512\rho\log(T\frac{\Delta_i^4}{16\rho^2})}{\Delta_i^2}\right)$$

C.3 Case b2:

In this case we will consider that the cluster containing the optimal arm a^* was eliminated by a sub-optimal cluster. Firstly, if conditions of case b1 holds then the optimal arm $a^* \in C_g$ will not be eliminated in round $g = g_*$ or it will lead to the contradiction that $r_i > r^*$ where $a_i, a^* \in C_g$. In any round g_*, if the optimal arm a^* gets eliminated then for any round from 1 to g_j all arms $a_j \in C_g$ such that $\sqrt{\rho\epsilon_g} < \frac{\Delta_j}{2}$ were eliminated according to assumption in case b1. Let, the arms surviving till g_* round be denoted by $C_g^{'}$. This leaves any arm a_b such that $\sqrt{\rho\epsilon_g} \geq \frac{\Delta_b}{2}$ to still survive and eliminate arm a^* in round g_*. Let, such arms that survive a^* belong to $C_g^{''}$. Also maximal regret per step after eliminating a^* is the maximal Δ_j among the remaining arms a_j with $g_j \geq g_*$. Let g_b be the round when $\sqrt{\rho\epsilon_g} < \frac{\Delta_b}{2}$ that is $g_b = min\{g|\sqrt{\rho\epsilon_g} < \frac{\Delta_b}{2}\}$. Hence, the maximal regret after eliminating the arm a^* is upper bounded by,

$\sum_{g_*=0}^{max_{j\in C_g^{'}} g_j} \sum_{i\in C_g^{''}:g_i>g_*} \left(\frac{2}{(T\epsilon_g^2)^\rho}\right).T \max_{j\in C_g^{''}:g_j\geq g_*} \Delta_j$

But, we know that for any round g, elements of C_g are the best performers in their respective clusters. So, taking that into account and $A' \supset C_g^{'}$ and $A'' \supset C_g^{''}$

$\leq \sum_{g_*=0}^{max_{j\in A'} g_j} \sum_{i\in A'':g_i>g_*} \left(\frac{2}{(T\epsilon_g^2)^\rho}\right).T \max_{j\in A'':g_j\geq g_*} \Delta_j$

$\leq \sum_{g_*=0}^{max_{j\in A'} g_j} \sum_{i\in A'':g_i>g_*} \left(\frac{2}{(T\epsilon_g^2)^\rho}\right).T.2\sqrt{\rho\epsilon_g}$, since $\sqrt{\rho\epsilon_g} < \frac{\Delta_i}{2}$

$\leq \sum_{g_*=0}^{max_{j\in A'} g_j} \sum_{i\in A'':g_i>g_*} \left(\frac{4T^{1-\rho}}{\epsilon_g^{2\rho-\frac{1}{2}}}\right)$

$\leq \sum_{i\in A'':g_i>g_*} \sum_{g_*=0}^{\min\{g_i,g_b\}} \left(\frac{4T^{1-\rho}}{2^{(2\rho-\frac{1}{2})g_*}}\right)$

$\leq \sum_{i\in A'} \left(\frac{4T^{1-\rho}}{2^{(2\rho-\frac{1}{2})g_*}}\right) + \sum_{i\in A''\setminus A'} \left(\frac{4T^{1-\rho}}{2^{(2\rho-\frac{1}{2})g_b}}\right)$

$\leq \sum_{i\in A'} \left(\frac{4T^{1-\rho} * 2^{2\rho-\frac{1}{2}}}{\Delta_i^{4\rho-1}}\right) + \sum_{i\in A''\setminus A'} \left(\frac{4T^{1-\rho} * 2^{2\rho-\frac{1}{2}}}{b^{4\rho-1}}\right)$

$\leq \sum_{i\in A'} \left(\frac{T^{1-\rho}2^{2\rho+\frac{3}{2}}}{\Delta_i^{4\rho-1}}\right) + \sum_{i\in A''\setminus A'} \left(\frac{T^{1-\rho}2^{2\rho+\frac{3}{2}}}{b^{4\rho-1}}\right)$

Summing up **Case a** and **Case b1** and **Case b2**, the total regret till round g is given by,

$R_T \leq \sum_{i\in A:\Delta_i\geq b} \left(\frac{2^{1+4\rho}\rho^{2\rho}T^{1-\rho}}{\Delta_i^{4\rho-1}}\right) + \left(\Delta_i + \frac{512\rho\log(T\frac{\Delta_i^4}{16\rho^2})}{\Delta_i}\right) + \left(\frac{T^{1-\rho}2^{2\rho+\frac{3}{2}}}{\Delta_i^{4\rho-1}}\right)\Big\} + \sum_{i\in A:0\leq\Delta_i\leq b} \left(\frac{T^{1-\rho}2^{2\rho+\frac{3}{2}}}{b^{4\rho-1}}\right) + max_{i:\Delta\leq b}\Delta_i T$

□

Again, we see that $\rho = 1$ brings out a near equivalent result as the result of proposition 5. That is, for $\rho = 1$,

$R_T \leq \sum_{i\in A:\Delta_i\geq b} \left\{\left(\frac{44}{(\Delta_i)^3}\right) + \left(\Delta_i + \frac{512\log(T\frac{\Delta_i^4}{16})}{\Delta_i}\right)\right\} + \sum_{i\in A:0\leq\Delta_i\leq b} \frac{12}{b^3} + max_{i:\Delta\leq b}\Delta_i T$

So, the most significant term for cluster elimination that is $\frac{512\log(T\frac{\Delta_i^4}{16})}{\Delta_i}$ is slightly more than arm elim-

ination most significant term $\frac{32\log(T\frac{\Delta_i^4}{16})}{\Delta_i}$ because of the constant, but its order is nearly same as arm elimination regret bound.

The principal takeaway from this result is that a lower value of $\rho \in (0,1]$ makes the algorithm risky(by increasing the error bound), but reduces expected regret whereas a higher value of $\rho > 1$ actually makes the algorithm less risky but at a cost of higher expected regret. The risk stems from the fact that the probability of sub-optimal arm elimination increases without proper exploration. So, there is always this trade-off between exploration and risk. So, in our algorithm we decrease the ρ in a graded fashion halving after every round but ρ is always bounded that is, $\rho \in (0,1)$.

D Appendix D

Theorem 2. *Considering both the arm elimination and cluster elimination condition, the total regret till* T *is upper bounded by* $R_T \leq \sum_{i\in A:\Delta\geq b}\left\{3\Delta_i + \left(\frac{44}{\Delta_i^3}\right) + \left(\frac{2^{1+4\rho}\rho^{2\rho}T^{1-\rho}}{\Delta_i^{4\rho-1}}\right) + \left(\frac{32\log(T\frac{\Delta_i^4}{16})}{\Delta_i}\right) + \left(\frac{512\rho\log(T\frac{\Delta_i^4}{16\rho^2})}{\Delta_i}\right)\right\} + \sum_{i\in A:0\leq\Delta_i\leq b}\left\{\left(\frac{12}{b^3}\right) + \left(\frac{T^{1-\rho}2^{2\rho+\frac{3}{2}}}{\Delta_i^3}\right) + \left(\frac{T^{1-\rho}2^{2\rho+\frac{3}{2}}}{b^{4\rho-1}}\right)\right\} + max_{i:\Delta\leq b}\Delta_i T$, *where* $\rho \in (0,1]$ *and* T *is the horizon.*

Proof. Combining both the cases of Proposition 5 and Proposition 6 we can see that a sub-optimal arm a_i can only be eliminated given that either m_i or g_i happens. In Proposition 5 we consider only arm elimination and in Proposition 6 we consider only cluster elimination. One vital point we point out is that, ϵ_m(in proposition 5) = ϵ_g(in proposition 6. Also we cluster the arms based on ϵ_m.

D.1 Case a:

So, we take the summation of the two events mentioned in Proposition 5(case a) and Proposition 6(case a) which gives us an upper bound on the regret given that the optimal arm a^* is still surviving,
$\leq \sum_{i\in A}\left\{\left(\frac{32}{\Delta_i^3}\right) + \left(\frac{2^{1+4\rho}\rho^{2\rho}T^{1-\rho}}{\Delta_i^{4\rho-1}}\right)\right\}$
before a sub-optimal arm is eliminated by arm elimination or cluster elimination condition.

D.2 Case b1:

Again, combining Proposition 5(case b1) and Proposition 6(case b1), we can show that till an arm or a cluster is eliminated, the maximum regret suffered due to pulling of a sub-optimal arm(or a sub-optimal cluster) is no less than,

$\sum_{i\in A}\left\{\left(\Delta_i + \frac{32\log(T\frac{\Delta_i^4}{16})}{\Delta_i}\right) + \left(\Delta_i + \frac{512\rho\log(T\frac{\Delta_i^4}{16\rho^2})}{\Delta_i}\right)\right\}$

D.3 Case b2:

Lastly we have to take into consideration the error bound, that the optimal arm a^* or the optimal cluster(that is the cluster containing the arm a^*) gets eliminated. Combining Proposition 5(case b2) and Proposition

6(case b2), we can show,

$\leq \sum_{i\in A'}\left(\frac{12}{\Delta_i^3}\right)+\sum_{i\in A''\setminus A'}\left(\frac{12}{b^3}\right)+\sum_{i\in A'}\left(\frac{T^{1-\rho}2^{2\rho+\frac{3}{2}}}{\Delta_i^3}\right)+\sum_{i\in A''\setminus A'}\left(\frac{T^{1-\rho}2^{2\rho+\frac{3}{2}}}{b^{4\rho-1}}\right)$

Hence, the total regret by combining **case a**, **case b1** and **case b2** is given by,

$$R_T \leq \sum_{i\in A:\Delta\geq b}\left\{\left(\frac{32}{\Delta_i^3}\right)+\left(\frac{2^{1+4\rho}\rho^{2\rho}T^{1-\rho}}{\Delta_i^{4\rho-1}}\right)+\left(\Delta_i+\frac{32\log\left(T\frac{\Delta_i^4}{16}\right)}{\Delta_i}\right)+\left(\Delta_i+\frac{512\rho\log\left(T\frac{\Delta_i^4}{16\rho^2}\right)}{\Delta_i}\right)\right\}+$$
$$\sum_{i\in A:0\leq\Delta_i\leq b}\left\{\left(\frac{12}{\Delta_i^3}\right)+\left(\frac{12}{b^3}\right)+\left(\frac{T^{1-\rho}2^{2\rho+\frac{3}{2}}}{\Delta_i^3}\right)+\left(\frac{T^{1-\rho}2^{2\rho+\frac{3}{2}}}{b^{4\rho-1}}\right)\right\}+max_{i:\Delta\leq b}\Delta_i T$$

$$R_T = \sum_{i\in A:\Delta\geq b}\left\{2\Delta_i+\left(\frac{44}{\Delta_i^3}\right)+\left(\frac{2^{1+4\rho}\rho^{2\rho}T^{1-\rho}}{\Delta_i^{4\rho-1}}\right)+\left(\frac{32\log\left(T\frac{\Delta_i^4}{16}\right)}{\Delta_i}\right)+\left(\frac{512\rho\log\left(T\frac{\Delta_i^4}{16\rho^2}\right)}{\Delta_i}\right)\right\}+$$
$$\sum_{i\in A:0\leq\Delta_i\leq b}\left\{\left(\frac{12}{b^3}\right)+\left(\frac{T^{1-\rho}2^{2\rho+\frac{3}{2}}}{\Delta_i^3}\right)+\left(\frac{T^{1-\rho}2^{2\rho+\frac{3}{2}}}{b^{4\rho-1}}\right)\right\}+max_{i:\Delta\leq b}\Delta_i T$$

Considering the case that all the arms are pulled once at the start,

$$R_T \leq \sum_{i\in A:\Delta\geq b}\left\{3\Delta_i+\left(\frac{44}{\Delta_i^3}\right)+\left(\frac{2^{1+4\rho}\rho^{2\rho}T^{1-\rho}}{\Delta_i^{4\rho-1}}\right)+\left(\frac{32\log\left(T\frac{\Delta_i^4}{16}\right)}{\Delta_i}\right)+\left(\frac{512\rho\log\left(T\frac{\Delta_i^4}{16\rho^2}\right)}{\Delta_i}\right)\right\}+$$
$$\sum_{i\in A:0\leq\Delta_i\leq b}\left\{\left(\frac{12}{b^3}\right)+\left(\frac{T^{1-\rho}2^{2\rho+\frac{3}{2}}}{\Delta_i^3}\right)+\left(\frac{T^{1-\rho}2^{2\rho+\frac{3}{2}}}{b^{4\rho-1}}\right)\right\}+max_{i:\Delta\leq b}\Delta_i T \qquad \square$$

E Appendix E

Algorithm	Cumulative Regret Upper Bound
UCB1	$O\left(\frac{K\log T}{\Delta}\right)$
UCB2	$O\left(K\left(\frac{(1+\epsilon(\alpha))log(T)}{2\Delta}+C(\alpha)\right)\right), 0<\alpha<1$
ϵ_n-greedy	$O\left(\frac{K\Delta\log T}{d^2}\right), 0<d<\Delta$
EXP3	$O\left(S\sqrt{KT\log(KT)}\right)$, where S is the hardness of the problem
UCB(δ)	$O\left(K\left(3\Delta+\frac{16}{\Delta}\log\left(\frac{2K}{\Delta\delta}\right)\right)\right)$, where δ is the error probability
UCB-Revisited	$O\left(\frac{K\log T\Delta^2}{\Delta}\right)$
MOSS	$\min\left\{O\left(\sqrt{KT}\right), O\left(\frac{K\log(T\Delta^2/K)}{\Delta}\right)\right\}$
KL-UCB	$O\left(K\left(\frac{\Delta\log(T)(1+\epsilon)}{d(r_i,r^*)}+\log(\log(T))+\frac{(\epsilon)}{T^{\beta(\epsilon)}}\right)\right)$, where $\epsilon>0$ and $d(r_i,r^*)>2\Delta_i^2$
UCB-Clustered	$O\left(\frac{K\log T\Delta^4}{\Delta}\right)$